informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 68, No. 4, July–August 2020, pp. 1218–1230
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Dynamic Scheduling of Multiclass Many-Server Queues with Abandonment: The Generalized $c\mu/h$ Rule

Zhenghua Long,[a] **Nahum Shimkin,**[b] **Hailun Zhang,**[c] **Jiheng Zhang**[d]

[a] School of Management, Nanjing University, Nanjing 210093, China; [b] Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel; [c] Institute for Data and Decision Analytics, The Chinese University of Hong Kong, Shenzhen, Shenzhen 518172, China; [d] Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Contact: zlong@nju.edu.cn, https://orcid.org/0000-0003-2925-375X (ZL); shimkin@ee.technion.ac.il, https://orcid.org/0000-0001-7105-9956 (NS); zhanghailun@cuhk.edu.cn, https://orcid.org/0000-0001-6116-6168 (HZ); jiheng@ust.hk, https://orcid.org/0000-0003-3025-1495 (JZ)

Received: September 8, 2018
Revised: May 24, 2019; June 26, 2019
Accepted: June 30, 2019
Published Online in Articles in Advance: May 22, 2020

Subject Classifications: queues: applications optimization, priority; healthcare
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2019.1908

Abstract. We study the fluid model of a many-server queue with multiple customer classes and obtain optimality results for this model. For the purpose of minimizing the long-run average queue-length costs and abandon penalties, we propose three scheduling policies to cope with any general cost functions and general patience-time distributions. First, we introduce the target-allocation policy, which assigns higher priority to customer classes with larger deviation from the desired allocation of the service capacity and prove its optimality for any general queue-length cost functions and patience-time distributions. The $Gc\mu/h$ rule, which extends the well-known $Gc\mu$ rule by taking abandonment into account, is shown to be optimal for the case of convex queue-length costs and nonincreasing hazard rates of patience. For the case of concave queue-length costs but nondecreasing hazard rates of patience, it is optimal to apply a fixed-priority policy, and a knapsack-like problem is developed to determine the optimal priority order efficiently. As a motivating example of the operations of emergency departments, a hybrid of the $Gc\mu/h$ rule and the fixed-priority policy is suggested to reduce crowding and queue abandonment. Numerical experiments show that this hybrid policy performs satisfactorily. We also prove the asymptotic optimality of policies in the original queueing system using the fluid results.

Funding: Z. Long's research is partially funded by the National Natural Science Foundation of China [Grant 71871114]. N. Shimkin's research is supported by the United States–Israel Binational Science Foundation [Grant 2006379]. H. Zhang's research is supported by the Shenzhen Research Institute of Big Data and Shenzhen Peacock Plan [Grant KQTD2015033114415450]. J. Zhang's research is supported in part by the Hong Kong Research Grants Council [GRF Grants 16201417 and 16501015].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2019.1908.

Keywords: multiclass queue • abandonment • fluid model • dynamic scheduling • knapsack problem

1. Introduction

In 2011, the number of left-without-being-seen (LWBS) patients in the United States was 2.6 million (The National Hospital Ambulatory Medical Care Survey) for the most common reason of being "fed up with waiting" (Rowe et al. 2006). Patient crowding in the emergency department (ED) has become an increasing public health problem for hospitals around the world, as it contributes to increased LWBS rates and dissatisfaction with care (Pines et al. 2011). We consider the problem of scheduling triage patients from the waiting room to treatment rooms to reduce ED crowding and LWBS rates.

Upon arrival, patients are rapidly sorted into five triage classes by experienced triage nurses using the Emergency Severity Index (ESI). The acuity levels from level 1 (most critical) to 5 (least critical) are based on patient acuity and resource needs (Gilboy et al. 2011). The ESI may or may not lead to improved patient flow through the ED because the physician response times for levels 1 and 2 are within minutes, but leaves the majority of lower-acuity patients waiting to be called for service according to their triage levels. Many patients visiting EDs are in low-acuity conditions. These patients have limited patience and may abandon the ED before receiving treatment. A new empirical study (Batt and Terwiesch 2015) indicates that the proportion of patients who abandon is up to 6.5%, and this rate ranges from 1.5% to 9.0% for different triage levels. The fundamental question that ED physicians face on a daily basis is: Which patient should be called for service first when a treatment bed becomes available? This also gives us a motivating example for treating a general queueing control problem—scheduling of multiclass many-server queues with abandonment.

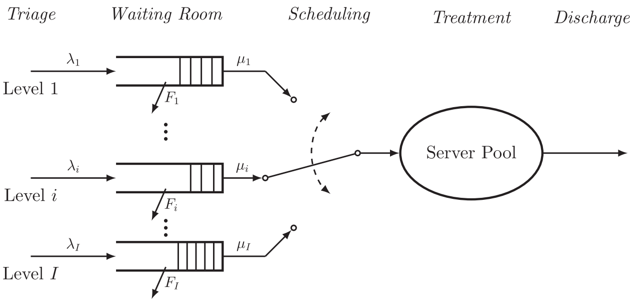
Recent studies on this scheduling problem have introduced a handy policy—namely, the $c\mu/\theta$ rule. This fixed-priority scheduling policy has been proved to be asymptotically optimal (Atar et al. 2008, 2010, 2011, 2014) for linear costs and exponential patience. It is consistent with the ESI system in the sense that high-acuity patients receive high priority. However, this rough treatment ignores the real-time status of the ED system and may lead to long waiting times and high LWBS rates for low-acuity patients. Indeed, the well-known generalized $c\mu$ rule ($Gc\mu$) assigns dynamic priority to the flows of multiple classes of customers (van Mieghem 1995, Mandelbaum and Stolyar 2004, Gurvich and Whitt 2009b). Recently, this scheduling policy has been applied in the control of patient flows in EDs with feedback (Huang et al. 2015). However, the $Gc\mu$ rule does not consider the LWBS patients. In this paper, we take into account patience time (the amount of time a patient is willing to wait for service) following general distributions. A natural paradigm to study the ED dynamics would be a multiclass, many-server queueing system with abandonment (the LWBS phenomenon), as shown in Figure 1. One of our main results is to introduce a dynamic scheduling policy, which we refer to as the generalized $c\mu/h$ rule ($Gc\mu/h$), to minimize the long-run average queueing costs and abandon penalties.

To describe our $Gc\mu/h$ rule, let μ_i be the service rate of level-i patients and F_i denote the patience-time distribution of level-i patients with the hazard-rate function h_i. Denote the marginal queue-length cost function and the penalty for each abandonment of level i by $c_i(\cdot)$ and γ_i, respectively. The arrival rates λ_i's are determined by triage nurses when categorizing ED visits. Let $B_i(t)$ be the number of level-i patients being served in the treatment rooms. We call the scheduling policy that serves the level-i patient [first-come-first-served (FCFS) within each level] with the highest index

$$i \in \arg\max_i \left(\frac{c_i\left(\lambda_i \int_0^{F_i^{-1}(1-B_i(t)\mu_i/\lambda_i)} F_i^c(s)ds\right)\mu_i}{h_i\left(F_i^{-1}\left(1 - B_i(t)\mu_i/\lambda_i\right)\right)} + \gamma_i\mu_i \right),$$

the *generalized* $c\mu/h$ *rule* ($Gc\mu/h$). We will show that the $Gc\mu/h$ rule is asymptotically optimal for convex queueing costs and nonincreasing hazard rates.

Figure 1. The Scheduling Problem in EDs with LWBS Patients



The $Gc\mu/h$ rule can be brought into play in systems like EDs due to its flexibility. For call-center operations, the latest information technology allows all agents and supervisors to observe the real-time status of the system (Gans et al. 2003). However, the situation in EDs is quite different. The queue status is usually unknown to ED staff because they are not notified when patients quit waiting. Our scheduling decision suitably depends on the current number of patients in the treatment room. Moreover, there is no need to modify the rule when the service capacity in the hospital changes. For example, the ED beds may be temporarily added to increase available capacity when all licensed beds are occupied (Derlet et al. 2014). In such a situation, the $Gc\mu/h$ rule adapts automatically to the change in service capacity.

Our $Gc\mu/h$ rule and the family of $Gc\mu$ rules (van Mieghem 1995, Mandelbaum and Stolyar 2004) all consider convex queue-length costs, but a theoretical understanding of more general cost functions is still lacking. To tackle this problem, we propose another dynamic scheduling policy referred to as the target-allocation policy (see Section 3.1). In an overcrowded ED, where a portion of the patients may end up leaving without being treated, the number of patients will be stable. The steady state of all types of patients in the treatment rooms can be viewed as an allocation of the service capacity. Our target-allocation policy aims to assign higher priority to the class of patients that deviates most from the optimal allocation, which is determined by solving a nonlinear optimization problem (13). The advantage of this policy is that it is asymptotically optimal for any general cost functions and patience distributions. However, the primary challenge lies in solving the nonlinear programming in advance.

The current practice in the EDs is mainly to implement triage priority (Batt and Terwiesch 2015), which can be considered as a fixed-priority policy. As mentioned in the above, the $Gc\mu/h$ rule (a dynamic-priority policy) is asymptotically optimal for convex queue-length costs and nonincreasing hazard-rate functions. Unexpectedly, for concave queue-length cost functions and nondecreasing hazard-rate functions of patience, we find that the optimal scheduling is a fixed-priority policy. In order to determine an optimal priority order, it involves the minimization of a concave function. As it is nontrivial to solve a concave optimization problem by using standard nonlinear approaches, we formulate it as a knapsack-like problem and develop a dynamic-programming algorithm. The algorithm can efficiently determine the treatment priority, especially when patients are further categorized by disease types. Our algorithm reduces the time complexity in a similar problem studied in Burke et al. (2008) (see Remark EC.1 in the e-companion). Until now, the three proposed policies actually allow us to choose the most appropriate policy for any given queue-length cost functions and patience distributions.

1.1. Literature Review

Fluid approximations for many-server queues with general patience-time distributions began to emerge following the pioneering work of Whitt (2006). Bassamboo and Randhawa (2010) established the optimal gap of fluid approximation as the system size increases. As an example of how powerful the fluid model approach is that it can be used to approximate a system with dependent service and patience times (see Bassamboo and Randhawa 2016, Wu et al. 2019). For multiclass queues, Atar et al. (2014) established the fluid limit of a multiclass $G/GI/n + GI$ queueing system, building on the approach developed by Kaspi and Ramanan (2011). Our fluid model is tailored to a multiclass $G/M/n + GI$ system with exponential service-time distributions.

The $c\mu$-type rules have a long history in the study of scheduling problems. As early as Smith (1956) and Cox and Smith (1961), the $c\mu$ rule was proposed and proved to be optimal for a multiclass $M/G/1$ system with linear holding costs. Recently, in Atar et al. (2008, 2010, 2011, 2014), it was extended to the $c\mu/\theta$ rule that is asymptotically optimal for a multiclass many-server queueing system with exponential patience and linear holding costs. The $Gc\mu$ rule of van Mieghem (1995) appears to be the first to consider nonlinear, convex holding costs in the analysis of a multiclass $G/G/1$ queue. Mandelbaum and Stolyar (2004) generalized the $Gc\mu$ rule to a system with heterogeneous servers. Our $Gc\mu/h$ rule extends van Mieghem (1995) and Atar et al. (2008, 2010, 2011, 2014) to a multiclass many-server queueing system with general patience and nonlinear holding costs.

Other than the $c\mu$-type rules, there has also been an expanding body of literature on the optimal control of multiclass queueing systems. Harrison and López (1999) explicitly solved a dynamic control problem in the multiclass parallel-server setting. Based on the conventional heavy-traffic regime, Ata and Tongarlak (2013) and Kim and Ward (2013) considered dynamic policies by studying the approximating Brownian control problems. Focusing on the Halfin–Whitt scaling proposed by Halfin and Whitt (1981) in the quality-and-efficiency-driven regime, Atar et al. (2004), Atar (2005), and Ata and Gurvich (2012) studied dynamic scheduling policies by formulating a Hamilton–Jacobi–Bellman equation based on the heavy traffic limits; Dai and Tezcan (2008) developed robust control policies to minimize the total linear holding and abandon costs for a parallel server system; Gurvich and Whitt (2009a, b, 2010) studied the staffing and control problems of service systems with multiple customer classes and multiple agent pools; and Kim et al. (2018) solved a diffusion-control problem to propose a scheduling policy for a critically loaded multiclass system with abandonment.

1.2. Contributions

The main contributions of this paper are summarized as follows:

- We propose three scheduling policies to control a multiclass many-server queueing system with all kinds of queue-length cost functions and patience distributions. The asymptotic optimality of the proposed policies is proved based on the results of the fluid model.
- The target-allocation policy is asymptotically optimal for any general queue-length cost functions and patience-time distributions by assigning higher priority to customer classes that deviate most from the desired allocation of the service capacity.
- The $Gc\mu/h$ rule extends the $Gc\mu$ rule of van Mieghem (1995) to overloaded systems with impatient customers and is shown to be asymptotically optimal for convex queue-length cost functions and nonincreasing hazard rates of patience.
- The fixed-priority policy is proved to be asymptotically optimal for concave queue-length cost functions and nondecreasing hazard rates of patience. It represents a generalization of the $c\mu/\theta$ rule of Atar et al. (2008, 2010, 2011, 2014), which considers linear cost and exponential patience.

The remainder of this paper is organized as follows. In Section 2, we introduce the fluid model of a multiclass many-server queueing system with abandonment (the original queueing system is analyzed in Section EC.2 of the e-companion). We also study a steady-state optimization problem. Our proposed policies and the main results are presented in Section 3. In Section 4, we use simulation experiments to test the performance of a hybrid policy. We show the connection between queueing and knapsack problems in Section 5. Our conclusion is stated in Section 6. Technical proofs and the analysis of the original queueing system are collected in the e-companion, where we also develop a dynamic-programming algorithm to solve the knapsack problem.

2. Multiclass Many-Server Queues

We consider the scheduling problem of a $G/M/n + GI$ queueing system with multiple customer classes. The system consists of n homogeneous servers that serve I classes of customers. Upon arrival, if a customer cannot be served immediately, this customer will be queued in a buffer. Each class-i customer has an independent patience time following distribution F_i for waiting in queue and abandons the queue once the waiting time exceeds the patience time. Within each class, customers are sent to servers according to the first-come-first-served discipline. Once admitted to service, a class-i customer will be served with exponentially distributed service time with mean $1/\mu_i$. Note that in the ED context, customer classes are usually called acuity levels; hereafter, we use these terms interchangeably. Such a system has been studied in Atar et al. (2008, 2010, 2011, 2014) under a fixed-priority policy with linear queue-length costs. The main difference is that our paper proposes three dynamic-priority policies in accordance with more general cost functions. As the stochastic system is analogous to that of Atar et al. (2014), the analysis of the original queueing model (including the asymptotic analysis of the fluid-scaled stochastic processes) will be placed in the e-companion (see Section EC.2). The main body of this paper will focus on the analysis of the fluid model.

2.1. A Fluid Model

The fluid model consists of I classes of fluid content that arrives at a service system having I unlimited waiting queues and a server pool with a fixed service capacity $n > 0$. Here, the stochastic counterpart of the fluid content is just the customers in the original queueing system. For each class $i = 1, \ldots, I$, the amount of external arrivals over $[0, t]$ is $E_i(t) = \lambda_i t$, where $\lambda_i > 0$. At time t, the arrival enters the server pool if there is any available service resource. Otherwise, the arrivals that cannot be directly served will join the end of their own queue and are allowed to abandon the queue once losing patience. We use $Q_i(t)$ and $B_i(t)$ to denote the amount of class-i fluid content waiting in queue and being served in the server pool, respectively. Thus, the total amount of class-i fluid content in the system is $X_i(t) = Q_i(t) + B_i(t)$.

Let $K_i(t)$ denote the total amount of class-i fluid content that has entered service by time t and $D_i(t)$ be the total amount of class-i fluid content that has completed service by time t. It is clear that the cumulative processes $K_i(t)$ and $D_i(t)$ would be nondecreasing. We can also deduce the following balance equation for B_i:

$$B_i(t) = B_i(0) + K_i(t) - D_i(t). \tag{1}$$

Obviously, there is also

$$\sum_{i=1}^{I} B_i(t) \le n. \tag{2}$$

Let the service time follow the distribution function $G_i(x) = 1 - e^{-\mu_i x}$ for class-i fluid content—namely, the service rate of class i is μ_i. Because of the memoryless property of exponential distributions, the service-completion process satisfies the equation

$$D_i(t) = \mu_i \int_0^t B_i(s)ds. \tag{3}$$

One can see that the derivative of the service-completion process is $\mu_i B_i(t)$, which facilitates the analysis of the convergence of the fluid model.

Because of the general patience-time distributions, we use the fluid measure-valued process developed in Atar et al. (2014) to capture the dynamics of the queues. Let $\eta_{i,t}([0,x])$ denote the amount of class-i fluid that has not abandoned by time t with elapsed time since arrival not longer than x no matter whether the fluid content has entered service or not. Within each queue, the fluid content is served based on the FCFS discipline. Thus, the fluid queue-length process of class i can be recovered as

$$Q_i(t) = \eta_{i,t}([0, w_i(t)]), \tag{4}$$

where $w_i(t)$ is the waiting time of the fluid content at the head of the class-i queue. Let $R_i(t)$ be the total amount of class-i fluid that abandons the queue during the time interval $[0, t]$. So, we have the following balance equation for Q_i:

$$Q_i(t) = Q_i(0) + E_i(t) - R_i(t) - K_i(t). \tag{5}$$

Let $F_i(\cdot)$ be the patience-time distribution of class-i fluid content. Then, we have

$$\eta_{i,t}([0,x]) = \int_{t-x}^{t} F_i^c(t-s)dE_i(s), \tag{6}$$

where $F_i^c(\cdot) = 1 - F_i(\cdot)$. Indeed, $dE_i(s)$ is the amount of fluid that enters the system at time s, among which $F_i^c(t-s)dE_i(s)$ is the amount that has not abandoned by time t. For $s < 0$, we regard $dE_i(s)$ as the fluid that had entered the system before time 0. On the other hand, $\eta_{i,t}([0,x])$ only consists of the arrivals between time $t - x$ to t. Thus, (6) holds. Clearly, $\eta_{i,t}(dx)$ is the density of class-i fluid with the waiting time x, but without abandoning at time t. Let the hazard-rate function of F_i be $h_i(x) = f_i(x)/F_i^c(x)$. Then, $h_i(x)$ is the fraction of the infinitesimal $\eta_{i,t}(dx)$ that abandons the queue. Recall that $w_i(t)$ is the longest elapsed time of the fluid in the class-i queue at time t, so the total amount of fluid that abandons the queue during the interval $[0, t]$ can be written as

$$R_i(t) = \int_0^t \left(\int_0^{w_i(s)} h_i(x)\eta_{i,s}(dx) \right) ds. \tag{7}$$

We denote by Π the class of all fluid work-conserving policies that, for all $t \ge 0$, satisfy

$$\left(n - \sum_{i=1}^{I} B_i(t) \right) \sum_{i=1}^{I} Q_i(t) = 0. \tag{8}$$

We refer to Equations (1)–(8) as the *fluid model* of a multiclass many-server queueing system. We rigorously prove in Theorem EC.1 in the e-companion that the tuple $(E, B, X, Q, D, K, R, \eta)$ satisfying (1)–(8) serves as the fluid limit of a multiclass many-server queueing system (see Section EC.2 of the e-companion for detailed discussion).

To manage such a system well, the cost it incurs should also be considered. We allow any general nondecreasing function $C_i(\cdot)$ for the (fluid) queue-length cost of each class i. Set $C_i(0) = 0$, which means there won't be any queue-length cost once there is no queue. There is also a penalty cost γ_i associated with abandonment for each class-i fluid content. Therefore, for any fluid work-conserving policy $\pi \in \Pi$, the average cost of the fluid model over $[0, T]$ is

$$J_T(\pi) = \frac{1}{T}\sum_{i=1}^{I}\left[\int_0^T C_i(Q_i(s))ds + \gamma_i R_i(T)\right]. \qquad (9)$$

The cost function of the original queueing system is defined in (EC.17) of the e-companion.

We define the traffic intensity as $\sum_{i=1}^{I} \lambda_i/\mu_i$. The system is underloaded if $\sum_{i=1}^{I} \lambda_i/\mu_i < n$, critically loaded if $\sum_{i=1}^{I} \lambda_i/\mu_i = n$, or overloaded if $\sum_{i=1}^{I} \lambda_i/\mu_i > n$. Intuitively, if the system is underloaded, then the average cost given above should vanish in the long run under any work-conserving policy. The following theorem validates this intuition.

Theorem 1. *If the system is underloaded—that is, $\sum_{i=1}^{I} \lambda_i/\mu_i < n$—then for any fluid work-conserving policy $\pi \in \Pi$, the fluid queue-length process of each class vanishes after a finite time, and the amount of fluid being served converges to λ_i/μ_i for each class $i = 1, \ldots, I$. As a consequence, the long-run average cost is zero. In other words, there exists a $T > 0$ such that $Q_i(t) = 0$ for all $t > T$,*

$$\lim_{t\to\infty} B_i(t) = \frac{\lambda_i}{\mu_i} \quad \text{and} \quad \lim_{T\to\infty} J_T(\pi) = 0.$$

The proof is postponed to Section EC.1 of the e-companion. A well-designed scheduling policy is expected to reduce system congestion, especially for an overloaded system. However, a critically loaded system also needs a well-designed scheduling policy. In Mandelbaum and Stolyar (2004), the $Gc\mu$ rule is applied to a queueing system with multiple types of customers and multiskilled servers. Note that their system is critically loaded, and the corresponding fluid model is studied under the $Gc\mu$ rule. We go one step further and focus on both the critically loaded and overloaded cases.

The following assumption on the input parameters is required throughout this paper.

Assumption 1 (On Input Parameters). *For each class $i = 1, \ldots, I$, the service-time distribution $G_i(x) = 1 - e^{-\mu_i x}$ is exponentially distributed, and the patience-time distribution $F_i(x) = \int_0^x f_i(y)dy$ is strictly increasing. The system is either critically loaded or overloaded—that is, $\sum_{i=1}^{I} \lambda_i/\mu_i \geq n$. The fluid queue-length cost function $C_i(\cdot)$ can be any differentiable nondecreasing function, and the marginal cost satisfies*

$$\frac{d}{dx}C_i(x) = c_i(x), \qquad (10)$$

where $c_i(x) \geq 0$. The abandon penalty cost also satisfies $\gamma_i \geq 0$.

Remark 1. It is well known that the steady-state behavior of the queue length of the fluid model of a single-class many-server queue depends upon the service-time distribution only through its mean, but upon the patience-time distribution beyond its mean (Whitt 2006). Therefore, we restrict ourselves to exponential service times. The simulation results in Section 4 suggest that our proposed policies also work well for nonexponential service times. However, for nonexponential service-time distributions, we are not able to prove that the fluid model converges to the invariant state, as time goes to infinity. But even for the single-class $G/GI/n + GI$ fluid model, this remains an open problem (see theorem 2 in Long and Zhang 2014, where an additional assumption on the initial state is needed for critically loaded and overloaded systems).

2.2. Stability and Optimality

We first give the following proposition to show the convergence relationship between the fluid content in the queues and that in service. This would help managers in scheduling the system when the status of the queues or the server pool cannot be fully observed. Usually the situation in waiting rooms in EDs is difficult to observe because the time when patients abandon the queue is normally not observed. This is one of the motivations for designing scheduling policies based on the status of the server pool in Section 3.

Proposition 1 (Equivalence of the Convergence of Q_i and B_i). *Given Assumption 1, for any fluid scheduling policy $\pi \in \Pi$, as $t \to \infty$,*

$$Q_i(t) \text{ converges} \Leftrightarrow B_i(t) \text{ converges} \quad \text{for all } i = 1, \ldots, I.$$

Moreover, for such a fluid-convergent policy, let F_i^{-1} be the inverse function of F_i. Then, we have, for all $i = 1, \ldots, I$,

$$q_i = \lambda_i \int_0^{F_i^{-1}\left(1 - b_i\mu_i/\lambda_i\right)} F_i^c(s)ds, \qquad (11)$$

where $q_i = \lim_{t\to\infty} Q_i(t)$ and $b_i = \lim_{t\to\infty} B_i(t)$, satisfying $0 \leq b_i \leq \lambda_i/\mu_i$ and $\sum_{i=1}^{I} b_i = n$. Therefore, $\lim_{T\to\infty} J_T(\pi) = \sum_{i=1}^{I} J_i(b_i)$. Here,

$$J_i(b_i) = C_i\left(\lambda_i \int_0^{F_i^{-1}\left(1 - b_i\mu_i/\lambda_i\right)} F_i^c(s)ds\right) + \gamma_i\left(\lambda_i - b_i\mu_i\right). \qquad (12)$$

The detailed proof is given in Section EC.1.2 of the e-companion. The steady-state behavior of the fluid content in the queues and of those being served follows the relation (11), which is consistent with theorem 3.1 in Whitt (2006). We can see from Proposition 1 that the steady-state behavior under the convergent policy has a simple form, and the cost function (12) can be expressed in terms of the status of the server pool.

Let us consider the optimization problem in terms of the steady state of the fluid model:

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{I} J_i(b_i) \\ &\text{subject to} \sum_{i=1}^{I} b_i \le n, \qquad (13) \\ &\qquad 0 \le b_i \le \frac{\lambda_i}{\mu_i}, i = 1, \dots, I. \end{aligned}$$

The decision variables b_i's can be intuitively understood as the amount of service resources that is assigned to class-i fluid content in the long run. The objective is to minimize the long-run average cost by choosing appropriate b_i's. The first constraint states that b_i's must be chosen so that the amount of fluid being served does not exceed the service capacity n. The second constraint implies that at most λ_i/μ_i service resource is needed to handle class i. Denote by $b^* = (b_1^*, \dots, b_I^*)$ an optimal solution to this nonlinear programming and J^* the optimal value. It is clear that b^* indicates the optimal allocation of the service capacity. Meanwhile, Proposition 1 implies that J^* is the lower bound of any fluid-convergent policies. The main goal of this paper is to find a scheduling policy that attains the lower bound.

Definition 1 (Stationary Optimal Control). A fluid-scheduling policy $\pi \in \Pi$ is said to be *stationary optimal* if the corresponding cost function (9) satisfies $\lim_{T\to\infty} J_T(\pi) = J^*$.

The following lemma implies that (13) can actually become either a convex or a concave optimization problem.

Lemma 1. *If the fluid queue-length cost functions C_i's are convex and the hazard-rate functions h_i's are nonincreasing, then the nonlinear programming* (13) *is a convex optimization problem. In contrast, if the fluid queue-length cost functions C_i's are concave and the hazard-rate functions h_i's are nondecreasing, then the nonlinear programming* (13) *is a concave optimization problem.*

A direct way to show the above lemma is to consider the derivative of the cost function $J_i(b_i)$. By (12) and after some basic calculations, it becomes clear that

$$\frac{d}{db_i} J_i(b_i) = -\frac{c_i\left(\lambda_i \int_0^{F_i^{-1}(1-b_i\mu_i/\lambda_i)} F_i^c(s)ds\right)\mu_i}{h_i\left(F_i^{-1}(1-b_i\mu_i/\lambda_i)\right)} - \gamma_i\mu_i. \qquad (14)$$

We leave the detailed proof to Section EC.1 of the e-companion. In the following section, we propose different scheduling policies for all types of optimization problems, such that the optimal value J^* can be attained in all cases.

3. Fluid Scheduling Policies

In this section, we propose fluid dynamic-priority policies that give a time-varying priority order. The goal is to design a policy such that the cost function (9) approaches J^*. In Section 3.1, the target-allocation policy is proposed for general queue-length cost functions and patience-time distributions. We then propose in Section 3.2 the $Gc\mu/h$ rule, which is an extension to the $Gc\mu$ rule in van Mieghem (1995), by adding abandonments. When the optimization problem (13) is convex, the $Gc\mu/h$ rule is shown to be stationary optimal. On the other hand, if (13) is a concave optimization problem, we find that it is optimal to apply the fixed-priority policy in Section 3.3.

Actually, every process in the fluid model has a stochastic counterpart. Therefore, our proposed policies can be easily translated back to the original queueing system. To be more rigorous, we define the stochastic version of our proposed polices in Section EC.2 of the e-companion. And we prove in Theorem EC.2 of the e-companion that the fluid-scaled queueing system under the stochastic version of the cost function (9) can also achieve the optimal value J^* of the nonlinear programming (13) asymptotically. Here, we stick to the design of fluid scheduling policies that help us better understand the original queueing system.

And so, we first introduce the *fluid dynamic-priority policy*. At time t, given that there is a certain amount of service resource, the policy chooses some amount of fluid content from the class with index

$$i \in \arg\max_{i=1,\dots,I} P_i(t), \qquad (15)$$

where $P_i(t)$ is the *priority value* for class i at time t. If the classes of fluid content with the highest priority value are all in service, then the available service resource can be assigned to classes with the second highest priority value, and so on and so forth. From this point of view, any (fluid) scheduling policy can be regarded as a (fluid) dynamic-priority policy. Equivalently, the fluid dynamic-priority policy means that the fluid content with lower priority can enter service at time t only if at that time no one else in the queue has higher priority. Therefore, the fluid dynamic-priority policy can also be expressed as

$$\int_0^t \sum_{\{j=1,\dots,I:P_j(s)>P_i(s)\}} Q_j(s)dK_i(s) = 0, \quad i = 1, \dots, I. \qquad (16)$$

Note that $\sum_{\{j=1,\dots,I:P_j(s)>P_i(s)\}} Q_j(s) = 0$ if $\{j = 1\dots, I : P_j(s) > P_i(s)\} = \emptyset$. As a special case, the (fluid) dynamic-priority policy becomes the *fixed-priority policy* when $P_i(t)$'s are independent of time t. We will see in Section 3.3 that (16) is actually an extension of equation (32) in Atar et al. (2014).

3.1. Target-Allocation Policy

We propose in this subsection a policy that is suitable for any general queue-length cost function and patience-time distribution. The optimal solution $b^* = (b_1^*, \dots, b_I^*)$ of (13) reveals that class-i fluid content should be allocated b_i^* amount of service resources in the long run. Thus, we define the following priority-value function:

$$P_i(t) = b_i^* - B_i(t), \tag{17}$$

for all $i = 1, \dots, I$. Intuitively, given the above priority-value function, the dynamic-priority policy serves the class with the largest deviation from its target. Thus, more service resources will be assigned to those classes that are not given enough service resources. All the B_i's will gradually be close to the optimal allocation b^* of the service capacity. We refer to this fluid-scheduling policy as the *target-allocation policy* denoted by π_{b^*} (see (EC.20) in the e-companion for the stochastic version). Its optimality is shown in Theorem 2 below, which is proved in Section EC.3.2 of the e-companion.

Theorem 2 (Optimality of the Target-Allocation Policy)**.** *Given Assumption 1, the fluid model (1)–(8) under the target-allocation policy π_{b^*} with the priority-value function (17) satisfies $\lim_{T\to\infty} J_T(\pi_{b^*}) = J^*$.*

3.2. The Generalized $c\mu/h$ Rule

For convex queue-length cost functions and patience-time distributions with nonincreasing hazard-rate functions under which the nonlinear programming (13) becomes a convex optimization by Lemma 1, we propose another dynamic-priority policy that is easier to implement. Consider the Lagrangian function

$$L(b_i, \alpha_0, \alpha_i, \beta_i) = \sum_{i=1}^{I} J_i(b_i) - \alpha_0\left(n - \sum_{i=1}^{I} b_i\right) - \sum_{i=1}^{I} \alpha_i b_i \mu_i - \sum_{i=1}^{I} \beta_i \cdot (\lambda_i - b_i\mu_i).$$

Combining it with (14), the optimal solution $b^* = (b_1^*, \dots, b_I^*)$ of (13) solves

$$\frac{c_i\left(\lambda_i \int_0^{F_i^{-1}(1-b_i^*\mu_i/\lambda_i)} F_i^c(s)ds\right)\mu_i}{h_i\left(F_i^{-1}(1-b_i^*\mu_i/\lambda_i)\right)} + \gamma_i\mu_i + \alpha_i\mu_i - \beta_i\mu_i = \alpha_0,$$
$$\alpha_i b_i^* = 0,$$
$$\beta_i \cdot (\lambda_i - b_i^*\mu_i) = 0,$$
$$\sum_{i=1}^{I} b_i^* = n,$$

where the Lagrange multipliers satisfy $\alpha_0 \in \mathbb{R}$ and $\alpha_i, \beta_i \geq 0$ for all $i = 1, \dots, I$. We assume that the cost function C_i, $i = 1 \dots, I$, satisfies conditions that are analogous to van Mieghem (1995, assumption 3) and Huang et al. (2015, assumption 2). Specifically, we have the following assumption.

Assumption 2 (Cost Regularity)**.** *The cost function C_i, $i = 1, \dots, I$ is strictly convex, and there is an interior solution to the minimization problem (13).*

Recall that the patience-time distribution F_i is strictly increasing. By Lemma 1, there is a unique solution to (13) if the cost functions are strictly convex and the hazard rates of patience are nonincreasing. If we assume in addition that $c_i(0) = 0$ and $\gamma_i = 0$, then all classes satisfy $b_i^* < \lambda_i/\mu_i$, making $\beta_i = 0$ for all i. Similarly, if we further assume that $h_i(x) \to 0$ as $x \to \infty$, then all classes receive positive service resources, making $\alpha_i = 0$ for all i. This essentially provides a sufficient condition such that the solution b_i^* is unique and interior.

Under Assumption 2, the Karush–Kuhn–Tucker (KKT) conditions then reduce to

$$\frac{c_i\left(\lambda_i \int_0^{F_i^{-1}(1-b_i^*\mu_i/\lambda_i)} F_i^c(s)ds\right)\mu_i}{h_i\left(F_i^{-1}(1-b_i^*\mu_i/\lambda_i)\right)} + \gamma_i\mu_i = \alpha_0, \tag{18}$$

$$\sum_{i=1}^{I} b_i^* = n. \tag{19}$$

Observe that the left-hand side of (18) is equal to a constant. This inspires us to consider the following priority-value function:

$$P_i(t) = \frac{c_i\left(\lambda_i \int_0^{F_i^{-1}(1-B_i(t)\mu_i/\lambda_i)} F_i^c(s)ds\right)\mu_i}{h_i\left(F_i^{-1}(1-B_i(t)\mu_i/\lambda_i)\right)} + \gamma_i\mu_i, \tag{20}$$

for all $i = 1, \dots, I$. This equation is referred to as the priority-value function of the *generalized $c\mu/h$ rule* ($Gc\mu/h$) denoted by π_G (see (EC.21) in the e-companion for the stochastic version).

The idea of the $Gc\mu/h$ rule comes from van Mieghem (1995), where the striking result $Gc\mu$ rule performs well for a single-server multiclass queueing system. Actually, Figure 1 in this paper is almost the same as figure 1 in van Mieghem (1995). The main difference is that our scheduling problem allows abandonment and considers a many-server pool. Later, the $Gc\mu$ rule was generalized to a system with heterogeneous servers in Mandelbaum and Stolyar (2004). They both consider the conventional diffusion approximation for critically loaded queueing systems without abandonment. We focus on the fluid model of an overloaded multiclass many-server queueing system with abandonment. This is why the hazard-rate function appears in the priority-value function (20). Another main difference is that we take advantage of the equivalence of the convergence of Q_i and B_i (see Proposition 1) to control the system based on the real-time value of $B_i(t)$ instead of $Q_i(t)$. The optimality of our $Gc\mu/h$ rule is shown in the following theorem, which we prove in Section EC.3.2 of the e-companion.

Theorem 3 (Optimality of the $Gc\mu/h$ Rule)**.** *Given Assumptions 1 and 2, if c_i and h_i are differentiable and the hazard-rate functions h_i's are nonincreasing, then the fluid model (1)–(8) under the $Gc\mu/h$ rule π_G with the priority-value function (20) satisfies $\lim_{T\to\infty} J_T(\pi_G) = J^*$.*

The assumption that c_i and h_i are differentiable is in the same spirit as the twice differentiability of C_i in section 4 of Mandelbaum and Stolyar (2004). It surprised us somewhat that the proofs of the optimality of the target-allocation policy and the $Gc\mu/h$ rule are almost the same. Part of the reason is that the priority-value functions go to a constant under both policies—the priority value of the target-allocation policy converges to 0, and that of the $Gc\mu/h$ rule converges to α_0. Therefore, we will prove Theorems 2 and 3 in Section EC.3.2 of the e-companion simultaneously.

3.3. Fixed-Priority Policy

A fixed-priority policy essentially prevents the fluid content from entering service as long as other classes of fluid content with higher priority are still waiting for their turn. Consider a priority order from class 1 (highest priority) to class I (lowest priority). Then, the priority-value function in (15) can be specified as

$$P_i(t) = I - i, \tag{21}$$

for all $i = 1, \ldots, I$. Note that only if the fluid content with the highest priority value are all in service, then the available service resource can be assigned to classes with the second highest priority value, and so on and so forth. Equation (16) becomes exactly the same as equation (32) in Atar et al. (2014). The following proposition shows that the system converges to the steady state under the fixed-priority policy (21). Especially, the limit of $B_i(t)$ follows the form as (23), which is the main feature of the fixed-priority policy. The proof is postponed to Section EC.3.3 of the e-companion.

Proposition 2 (Convergence of the Fixed-Priority Policy)**.** *Given Assumption 1, the fluid model (1)–(8) under the fixed-priority policy with the priority-value function (21) converges to the following steady state:*

$$\lim_{t\to\infty} B_i(t) = b_i \quad \textit{and} \quad \lim_{t\to\infty} Q_i(t) = q_i\,, \tag{22}$$

for all $i = 1, \ldots, I$, where the allocation $b = (b_1, \cdots, b_I)$ of the service capacity to their dedicated classes is

$$b = \left(\frac{\lambda_1}{\mu_1}, \cdots, \frac{\lambda_{i_0-1}}{\mu_{i_0-1}}, n - \sum_{j<i_0} \frac{\lambda_j}{\mu_j}, 0, \cdots, 0\right), \tag{23}$$

where $i_0 = \max\left\{i \in [1, \cdots, n] : \sum_{j=1}^{i-1} \frac{\lambda_j}{\mu_j} < n\right\}$. And

$$q_i = \begin{cases} 0, & i < i_0, \\ \lambda_i \displaystyle\int_0^{F_i^{-1}(1-b_i\mu_i/\lambda_i)} F_i^c(s)ds, & i = i_0, \\ \lambda_i \displaystyle\int_0^{\infty} F_i^c(s)ds, & i > i_0. \end{cases}$$

Moreover, there exists $T > 0$ such that $Q_i(t) = 0$ for all $t > T$ and $i = 1, \ldots, i_0 - 1$.

The allocation of the service capacity (23) takes a special form such that $b_i = \lambda_i/\mu_i$ for all classes $i < i_0$ being fully served, $b_i = 0$ for all classes $i > i_0$ without receiving any service, and $b_{i_0} = n - \sum_{i=1}^{i_0-1} \lambda_i/\mu_i$ for at most one class i_0 being partially served. This is virtually a solution on the boundary of the feasible region of (13). Therefore, if the nonlinear programming (13) is a concave optimization problem, then the optimal solution $b^* = (b_1^*, \ldots, b_I^*)$ surely has the same form as (23) after reordering the class indices if needed. This is associated with an optimal fixed-priority order, of which the corresponding fixed-priority policy is denoted by π_{P^*} (see (EC.22) in the e-companion for the stochastic version). Note that the order among the classes with $b_i^* = \frac{\lambda_i}{\mu_i}$ can be arbitrarily determined. It can also be arbitrary for those with $b_i^* = 0$.

Theorem 4 (Optimality of the Fixed-Priority Policy)**.** *Given Assumption 1, if the queue-length cost functions C_i's are concave and the hazard-rate functions h_i's are nondecreasing, then the fluid model (1)–(8) under the fixed-priority policy π_{P^*} with the priority-value function (21) (after reordering the class indices if needed) satisfies $\lim_{T\to\infty} J_T(\pi_{P^*}) = J^*$.*

Theorem 4 is proved in Section EC.3.3 of the e-companion. This theorem actually gives a sufficient condition for the optimality of the fixed-priority policy. We will show in Section 5 the innovative connection between the fixed-priority policy and knapsack problems.

Remark 2 (Connection to Linear Queue-Length Costs and Exponential Patience)**.** We consider a special case of exponential patience-time distributions $F_i(x) = 1 - e^{-\theta_i x}$ and linear queue-length cost functions by setting $C_i(x) = c_i x$ for all $i = 1, \ldots, I$. Then, the optimization problem (13) becomes the following linear programming:

$$\begin{aligned} &\text{minimize} \sum_{i=1}^{I} \left[c_i \frac{\lambda_i - \mu_i b_i}{\theta_i} + \gamma_i(\lambda_i - \mu_i b_i) \right] \\ &\text{subject to} \sum_{i=1}^{I} b_i \le n, \\ &\qquad 0 \le b_i \le \frac{\lambda_i}{\mu_i}, i = 1, \ldots, I. \end{aligned} \tag{24}$$

Let $\tilde{c}_i = c_i + \theta_i\gamma_i$ for notational simplicity. Then, the objective function in (24) is identical to

$$\text{maximize} \quad \sum_{i=1}^{I} \frac{\tilde{c}_i\mu_i}{\theta_i} b_i. \tag{25}$$

Because of the simple form of the above objective function, to maximize (25), the obvious solution is to assign as much value (namely, λ_i/μ_i) as possible to b_i with higher coefficient $\tilde{c}_i\mu_i/\theta_i$. For convenience, we relabel indices such that $\tilde{c}_1\mu_1/\theta_1 \geq \cdots \geq \tilde{c}_I\mu_I/\theta_I$. After reordering the indices, the linear programming (24) admits an optimal solution with the same form as (23). Thus, it is straightforward to design a fixed-priority policy that assigns higher priority to customers with higher $\tilde{c}_i\mu_i/\theta_i$. This is exactly the $c\mu/\theta$ rule studied in Atar et al. (2008, 2010, 2011, 2014). The optimality of the $c\mu/\theta$ rule can be easily seen from Propositions 1 and 2.

4. Numerical Experiments

We first introduce a hybrid policy that is a mixture of the fixed-priority policy and the $Gc\mu/h$ rule in Section 4.1. This policy can be implemented in EDs to reduce the crowding and LWBS rates. We illustrate with performance metrics including the numbers of patients in each of the five acuity levels in steady state and the long-run average cost that the hybrid policy inherits the merits of both the fixed-priority policy and the $Gc\mu/h$ rule. In Section 4.2, we present the parameters used in our experiments. Our simulation results in Section 4.3 show that the lengths of the queues for patients of levels 1 and 2 with the highest priority are close to zero in steady state. We also observe that the patients in the other three less-critical levels following the $Gc\mu/h$ rule are able to receive proper medical treatment in the long run.

4.1. A Hybrid Policy

In practice, we can combine the fixed-priority policy with the $Gc\mu/h$ rule. It is widely accepted that in EDs, patients are generally called for service on a FCFS basis by triage level (Batt and Terwiesch 2015). Actually, the $Gc\mu/h$ rule and the fixed-priority rule have their own merits in the sense that the former gives consideration to the least-critical patients, whereas the latter enables the most-critical patients to receive timely treatment. In view of the fact that the most-critical patients may not survive if they fail to receive medical care in time, there is no doubt that they should be given the highest priority. On the other hand, the majority of patients in low-acuity conditions should also be taken care of in a timely manner, as they are the main reason for ED crowding and high LWBS rates. To balance the tradeoff, we suggest a hybrid policy to improve patient flows in EDs as follows: According to ESI, assign the highest priority to level 1 and the second highest priority to level 2, and apply the $Gc\mu/h$ rule to levels 3, 4, and 5 with proper input parameters. The fluid queues of levels 1 and 2 will vanish after a finite time by Proposition 2. This means that all patients in levels 1 and 2 are prior to entering service and then all patients in levels 3, 4, and 5 will enter service according to the $Gc\mu/h$ rule. Then, by Theorem 3, the fluid model under the hybrid policy converges to a certain steady state.

4.2. Simulation Parameters

In order to demonstrate the fluid approximation, the service capacity is set to be $n = 100$. We now explain the parameters in Table 1. In the column titled "Arrival rate," we display the arrival rates λ_i's for different acuity levels. The service rates μ_i's are set to increase monotonically from level 1 to level 5, as is typically the case in EDs. In general, the monotonicity of the parameters in Table 1 is unnecessary. Because the hybrid policy assigns the highest priority to level 1 and the second highest priority to level 2, there is no need to identify the abandon penalty and queue-length cost for these two levels. An alternative way to think about this is that the cost of not treating the most critical patients promptly is high, and so they must be seen by a physician within minutes. We will see in the next subsection that there is almost no queue for level 1 and 2 patients. For level 3, 4, and 5 patients, the related costs are presented in the last two columns.

For patients in levels 1 and 2, we assume that they will not abandon the queue because of their high treatment priority. For patients in less critical conditions, their patience-time distributions are assumed to be $F_i(x) = 1 - 1/(x+1)$ for all levels $i = 3, 4, 5$, of which the hazard-rate function $h_i(x) = 1/(x+1)$ is

Table 1. Arrival and Service Rates Together with Related Costs for Five Triage Classes

Triage class	Arrival rate λ_i	Service rate μ_i	Abandon penalty γ_i	Queue-length cost $C_i(x)$
Level 1	30	1	—	—
Level 2	40	2	—	—
Level 3	80	3	3	$3x^2$
Level 4	100	4	2	$2x^2$
Level 5	160	5	1	x^2

nonincreasing. Considering the $Gc\mu/h$ rule for levels 3, 4, and 5 and applying the above parameters to (20) yield

$$P_i(t) = 2(6-i)\ln\left(\frac{\lambda_i}{B_i(t)\mu_i}\right)\frac{\lambda_i^2}{B_i(t)} + \gamma_i\mu_i \quad \text{for } i = 3,4,5. \tag{26}$$

Thus, once there are no more level 1 and 2 patients waiting, the patients in levels 3, 4, and 5 will be treated according to the above priority-value function.

Assume that the arrivals follow Erlang $E_2(1/\lambda_i)$ distributions for levels $i = 1,\ldots,5$. From now on, we use "$E_2(x)$" to denote an Erlang E_2 distribution with mean x, "expo(x)" to denote an exponential distribution with mean x, and "$\ln(x,y)$" to denote a log-normal distribution with mean x and variance y. As pointed out in Remark 1, the steady state of the fluid approximation depends only on the mean of the service-time distributions. Thus, we simulate the system with three different service-time distributions—that is, expo$(1/\mu_i)$, $E_2(1/\mu_i)$, and $\ln(1/\mu_i, 1/\mu_i^2)$—which have same service rate μ_i for any $i = 1,\ldots,5$.

With the given parameters and distributions, we run each simulation under the hybrid policy for 1,000 time units. The first 10% and the last 10% of the simulation period are regarded as the warm-up and the close-down periods of the system; thus, they are discarded when computing the steady-state performance metrics. We use the batch-means method with five independent runs to obtain confidence intervals.

4.3. Summary of Results

We present the results of our simulation experiments in this subsection. The steady state of the fluid model under the hybrid policy can be easily computed, given the experimental setting in Table 1 and the priority-value function (26). For level 1 and 2 patients with the highest priority, we can deduce from (23) that $b_1 = \frac{\lambda_1}{\mu_1} = 30$ and $b_2 = \frac{\lambda_2}{\mu_2} = 20$. Thus, the service capacity that remains for level 3, 4, and 5 patients is 50. And their steady state can be obtained by solving the KKT condition (18) with service capacity $b_3 + b_4 + b_5 = 50$. Then, the corresponding queue lengths q_i's, $i = 1,\ldots,5$, and the total cost follow directly from (11) and (12). This yields the fluid approximation of the system, which is displayed in the last column of Table 2 for comparison with the simulation results. In Table 2, we also present the simulation approximations for Q_i's, B_i's, and the total long-run average cost, along with their relative errors and 95% confidence intervals for three different service-time distributions. The relative errors for Q_1 and Q_2 are omitted because their fluid approximations are 0.

It is worth noting that the steady-state performance of the systems with general service times is similar to that of the system with exponential service-time distributions. For example, the value of B_3 is 15.758 when service-time distributions for different levels are exponential. The corresponding values of B_3 for Erlang E_2 and log-normal distributions are 15.730 and 15.711, respectively. The results of other performance metrics are also close to each other.

Table 2. Comparison of Simulation Results and Approximations with General Service-Time Distributions

	Exponential expo$(1/\mu_i)$		Erlang $E_2(1/\mu_i)$		Log-normal $\ln(1/\mu_i, 1/\mu_i^2)$		
Performance	Simulation	Relative error (%)	Simulation	Relative error (%)	Simulation	Relative error (%)	Approximation
Q_1	0.600 ±0.063	—	0.555 ±0.076	—	0.578 ±0.130	— —	0
Q_2	0.621 ±0.077	—	0.668 ±0.099	—	0.668 ±0.002	— —	0
Q_3	42.119 ±1.815	2.34	42.208 ±1.694	2.13	42.325 ±1.643	1.86	43.126
Q_4	49.865 ±1.847	0.91	49.783 ±1.929	1.07	49.816 ±1.904	1.01	50.325
Q_5	80.247 ±3.220	0.48	80.365 ±2.857	0.34	80.497 ±3.233	0.18	80.640
B_1	29.775 ±0.403	0.75	29.864 ±0.500	0.45	29.995 ±0.778	0.02	30
B_2	19.941 ±0.537	0.30	20.024 ±0.181	0.12	20.035 ±0.439	0.18	20
B_3	15.758 ±0.172	1.31	15.730 ±0.060	1.13	15.711 ±0.218	1.01	15.554
B_4	15.245 ±0.171	0.87	15.193 ±0.204	0.52	15.153 ±0.190	0.26	15.114
B_5	19.280 ±0.250	0.27	19.186 ±0.144	0.76	19.145 ±0.218	0.97	19.332
Long run average cost	18,027.311 ±562.222	3.66	17,833.704 ±414.350	2.55	18,050.739 ±556.930	3.80	17,390.018

Moreover, our approximations using the fluid steady state are fairly accurate. The relative errors of the approximations for Q_i's and B_i's are less than 2.34% and 1.31%, respectively, with an average error of 1.17% for patients who are waiting in queue and 0.59% for patients who are being treated. The quality of the approximations for the long-run average cost is relatively worse. Because of the quadratic queue-length cost functions in Table 1, the magnitude of the long-run average cost in the last row of Table 2 is much larger than that of the other performance metrics. Even so, the average error is still less than 3.34% across all simulations with different service-time distributions.

5. Knapsack Problems

In this section, we show the connection between queueing systems and knapsack problems. We declare that the $c\mu/\theta$ rule derived from (25) is identical to the Fractional Knapsack Problem (27). We also introduce the Fractional 0-1 Knapsack Problem in (28), which turns out to be consistent with the fixed-priority scheduling problem in Section 3.3. Moreover, in Section EC.4 of the e-companion, we propose a dynamic programming algorithm to solve it efficiently.

5.1. The Fractional Knapsack Problem

The *Fractional Knapsack Problem* (also known as the continuous knapsack problem) was first considered by George Dantzig in Dantzig (1957). Let there be K items, indexed by $k = 1, \ldots, K$, with value v_k and weight w_k for item k. This knapsack problem allows every item to be divided. The amount of item k that is packed in the knapsack will be denoted by y_k being a real number between 0 and w_i. The maximum weight that can be carried in the knapsack is W. More specifically, we wish to solve the following maximization problem:

$$\begin{aligned} &\text{maximize} \sum_{k=1}^{K} \frac{v_k}{w_k} y_k \\ &\text{subject to} \sum_{k=1}^{K} y_k \le W, \\ &\qquad 0 \le y_k \le w_k, k = 1, \ldots, K. \end{aligned} \tag{27}$$

Because of its very simple form, it admits an immediate algorithm: Order the items according to their value-to-weight ratio, $\frac{v_1}{w_1} \ge \cdots \ge \frac{v_K}{w_K}$, then apply a greedy algorithm to pack as many high-ratio items into the knapsack as possible. It can be easily seen that the form of the optimal solutions is either 0 or w_k for each item, with at most one exception to choose the fractional part of its weight. Now, comparing the maximization problems (25) and (27), there is no doubt that the $c\mu/\theta$ rule is virtually a Fractional Knapsack Problem. We formally state it in the following proposition and omit its proof for brevity.

Proposition 3. *For linear* queue-*length cost functions and exponential patience-time distributions, the $c\mu/\theta$ rule problem* (24) *is identical to the Fractional Knapsack Problem* (27).

5.2. The Fractional 0-1 Knapsack Problem

Instead of the linear objective functions in (27), we consider a nonlinear reward function $V_k(y_k)$ being the reward value of item k with weight y_k packed into the knapsack. For standardization, we set $V_k(0) = 0$. Also, $V_k(y_k)$ is postulated to be a nondecreasing function in y_k. Among all the possible choices of $\{y_1, y_2, \cdots, y_K\}$, we allow at most one item to be strictly between 0 and its maximum weight. Hence, the problem (27) is extended to

$$\begin{aligned} &\text{maximize} \sum_{k=1}^{K} V_k(y_k) \\ &\text{subject to} \sum_{k=1}^{K} y_k \le W, \\ &\qquad 0 \le y_k \le w_k, k = 1, \ldots, K, \\ &\qquad 0 < y_k < w_k \text{ for at most one } k \in \{1, \cdots, K\}. \end{aligned} \tag{28}$$

We refer to (28) as the *Fractional 0-1 Knapsack Problem* because it allows at most one item to be divided like in the Fractional Knapsack Problem and requires other items to be packed in their entirety or not packed at all like in the classical 0-1 Knapsack Problem. Obviously, the last constraint can be eliminated when (28) is a concave optimization problem. Now, it becomes clear that in order to find an optimal fixed-priority order, it is essential to solve the Fractional 0-1 Knapsack Problem. Therefore, the proposition below immediately follows.

Proposition 4. *For general queue-length cost functions and patience-time distributions, the fixed-priority control problem is equivalent to the Fractional* 0-1 *Knapsack Problem* (28).

Note that if we restrict ourselves to the family of fixed-priority policies, then there is no need to require the queue-length cost functions to be concave and the hazard rates to be nondecreasing, as in Theorem 4. All we need is to find an optimal solution on the boundary of the feasible region of (13) by adding a constraint like the last one in (28).

Remark 3. Note that in the study of knapsack problems, it is quite common to assume that all the weights are integer numbers—that is, W and w_k in (28) are all integers. It is also well known that the classical 0-1

Knapsack Problem can be solved in pseudo-polynomial time through dynamic programming (see, e.g., Martello and Toth 1990). In Section EC.4 of the e-companion, we develop a dynamic-programming algorithm to solve our fixed-priority control problem in the same manner, for which we need to assume that the related parameters—that is, λ_i and μ_i in (13)—are rational numbers.

6. Conclusion

To the best of our knowledge, this paper is the first to extend the $Gc\mu$ rule by adding abandonment with general patience-time distributions. We consider the control problem of a multiclass many-server queueing model with general holding cost functions and patience-time distributions based on the fluid approximation. To minimize the queue-length costs and abandon penalties, we solve a nonlinear programming in terms of the steady state of the fluid model. The optimal solution inspires us to design three fluid scheduling polices for the fluid model in Section 3. For the original queueing system, the stochastic version of the three scheduling polices is similarly defined in Section EC.2 of the e-companion. The target-allocation policy with the priority-value function (17) (see (EC.20) in the e-companion for its stochastic version) works for any kind of queue-length cost functions and patience-time distributions. Interestingly, we find that the $Gc\mu/h$ rule with the priority-value function (20) (see (EC.21) in the e-companion for its stochastic version) is asymptotically optimal for convex queue-length cost functions and nonincreasing hazard rates of patience. In contrast, the fixed-priority policy is asymptotically optimal for concave queue-length cost functions and nondecreasing hazard rates of patience with the priority-value function (21) (see (EC.22) in the e-companion for its stochastic version) after reordering the class indices, if needed. In order to find such an optimal order of indices, we develop a dynamic-programming algorithm (see Section EC.4 of the e-companion) based on the unexpected consistency between queueing and knapsack problems. Motivated by the application to EDs, a hybrid of the fixed-priority policy and the $Gc\mu/h$ rule is suggested to reduce patient abandonment and crowding in waiting rooms. The simulation results show that the performance of our proposed policy is fairly close to the theoretical result, with a relative error of less than 3.8% among all performance metrics.

Several extensions are possible for future research. First, we have assumed that the service-time distributions are exponential, which facilitates the equilibrium analysis of the fluid model. The corresponding convergence for the dynamically controlled multiclass many-server queue with nonexponential service-time distributions remains to be developed. Another direction is to develop priority-value functions based on the waiting time or the queue length. Although we believe that in EDs our proposed dynamic policies based on the number of patients being treated are more realistic, we could accommodate a wider range of situations if we were able to show the asymptotic optimality of a queue-length-based policy.

Acknowledgments

The authors thank the area editor, associate editor, and the anonymous reviewers for their careful reading of the paper and for providing constructive feedback. Useful discussions with Haya Kaspi and Avishai Mandelbaum at the initial part of this research are also gratefully acknowledged.

References

Ata B, Gurvich I (2012) On optimality gaps in the Halfin-Whitt regime. *Ann. Appl. Probab.* 22(1):407–455.

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

Atar R (2005) Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 15(4):2606–2650.

Atar R, Giat C, Shimkin N (2008) The $c\mu/\theta$ rule. Baras J, Courcoubetis C, eds. *Proc. 3rd Internat. Conf. Performance Evaluation Methodologies Tools, ValueTools '08* (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Brussels, Belgium), 58:1–58.4.

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many server queues with abandonment. *Oper. Res.* 58(5):1427–1439.

Atar R, Giat C, Shimkin N (2011) On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems*. 67(2):127–144.

Atar R, Kaspi H, Shimkin N (2014) Fluid limits for many-server systems with reneging under a priority policy. *Math. Oper. Res.* 39(3):672–696.

Atar R, Mandelbaum A, Reiman MI (2004) Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. *Ann. Appl. Probab.* 14(3):1084–1134.

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* 58(5):1398–1413.

Bassamboo A, Randhawa RS (2016) Scheduling homogeneous impatient customers. *Management Sci.* 62(7):2129–2147.

Batt RJ, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Sci.* 61(1):39–59.

Burke GJ, Geunes J, Edwin Romeijn H, Vakharia A (2008) Allocating procurement to capacitated suppliers with concave quantity discounts. *Oper. Res. Lett.* 36(1):103–109.

Cox D, Smith W (1961) *Queues*. Cox DR, Hinkley DV, Rubin D, Silverman BW, eds. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, vol. 2 (Taylor & Francis, Abingdon, UK).

Dai JG, Tezcan T (2008) Optimal control of parallel server systems with many servers in heavy traffic. *Queueing Systems* 59(2):95–134.

Dantzig GB (1957) Discrete-variable extremum problems. *Oper. Res.* 5(2):266–277.

Derlet RW, McNamara RM, Kazzi AA, Richards JR (2014) Emergency department crowding and loss of medical licensure: A new risk of patient care in hallways. *Western J. Emergency Medicine* 15(2):137–141.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Gilboy N, Tanabe T, Travers D, Rosenau AM (2011) *Emergency Severity Index (ESI): A Triage Tool for Emergency Departments* (Agency for Healthcare Research and Quality, Rockville, MD). Accessed September 17, 2019, http://www.ahrq.gov/professionals/systems/hospital/esi/esi1.html.

Gurvich I, Whitt W (2009a) Queue-and-idleness-ratio controls in many-server service systems. *Math. Oper. Res.* 34(2):363–396.

Gurvich I, Whitt W (2009b) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing Service Oper. Management* 11(2):237–253.

Gurvich I, Whitt W (2010) Service-level differentiation in many-server service system via queue-ratio routing. *Oper. Res.* 58(2):316–328.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.

Harrison JM, López MJ (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing Syst.* 33(4):339–368.

Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* 63(4):892–908.

Kaspi H, Ramanan K (2011) Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* 21(1):33–114.

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

Kim J, Randhawa RS, Ward AR (2018) Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manufacturing Service Oper. Management* 20(2):285–301.

Long Z, Zhang J (2014) Convergence to equilibrium states for fluid models of many-server queues with abandonment. *Oper. Res. Lett.* 42(6–7):388–393.

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

Martello S, Toth P (1990) *Knapsack Problems: Algorithms and Computer Implementations*, Wiley-Interscience Series in Discrete Mathematics and Optimization (John Wiley & Sons, New York).

Pines JM, Hilton JA, Weber EJ, Alkemade AJ, Al Shabanah H, Anderson PD, Bernhard M, et al (2011) International perspectives on emergency department crowding. *Acad. Emergency Medicine* 18(12):1358–1370.

Rowe BH, Channan P, Bullard M, Blitz S, Saunders LD, Rosychuk RJ, Lari H, Craig WR, Holroyd BR (2006) Characteristics of patients who leave emergency departments without being seen. *Acad. Emergency Medicine* 13(8):848–852.

Smith WE (1956) Various optimizers for single-stage production. *Naval Res. Logist. Quart.* 3(1-2):59–66.

van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Probab.* 5(3):809–833.

Whitt W (2006) Fluid models for multiserver queues with abandonments. *Oper. Res.* 54(1):37–54.

Wu CA, Bassamboo A, Perry O (2019) Service system with dependent service and patience times. *Management Sci.* 65(3):1151–1172.

Zhenghua Long is an assistant professor in management at the School of Management, Nanjing University. His research interests lie in asymptotic analysis and optimal control of queueing systems and their applications in manufacturing and services.

Nahum Shimkin is a professor and dean of the Viterbi Faculty of Electrical Engineering at the Technion. His research interests include stochastic control and planning, queueing systems, game theoretical analysis of multiuser systems, and reinforcement learning.

Hailun Zhang is an assistant professor in data and decision analytics at the Chinese University of Hong Kong, Shenzhen. His research interests include data-driven queueing networks, online algorithm design, and their applications.

Jiheng Zhang is an associate professor in industrial engineering and decision analytics at the Hong Kong University of Science and Technology. His research interests are in applied probability, stochastic modeling and optimization, data analysis, numerical methods, and algorithms.