TUTORIALS IN OPERATIONS RESEARCH
INFORMS 2019
This version: April 5, 2019

© 2019 INFORMS | ISBN 978-0-xxxxxxx-x-x
http://?

Chapter ??

Tutorial Paper: Scheduling an Overloaded Multiclass Many-Server Queue with Impatient Customers

Amber L. Puha*
California State University San Marcos, apuha@csusm.edu

Amy R. Ward[†]
University of Chicago, amy.ward@chicagobooth.edu

Abstract We describe a fluid model with time-varying input that approximates a multiclass many-server queue with time-varying arrivals (specifically, the multiclass $G/GI/N + GI$ queue). We show how to use the restricted fluid model with constant input rate to approximately solve scheduling control problems for a queue with constant arrival rate. The key is to characterize the invariant states of the fluid model, because they typically provide an approximation to the mean steady-state behavior of the queue under a wide range of scheduling policies. The resulting fluid control problem motivates using a static priority scheduling policy when the objective is to minimize the long run average abandonment rate, but may motivate a different class of scheduling policies when there are also holding costs. We end by discussing several open problems.

Keywords Scheduling, Many-server queue, Impatience, Abandonment, Reneging, Fluid Model

1. Introduction

Throughout a long history in the academic literature, scheduling problems have been studied by researchers in the fields of business, engineering, and mathematics. Scheduling attracts wide interest because of the central role it plays in many different application environments, including manufacturing and production systems [58], large-scale computing systems [32], service systems such as call centers [26, 1], and healthcare systems [35]. Fundamentally, scheduling problems ask how to pair incoming requests for service or processing with the resources available (whether they be machines or human employees). Scheduling endures as an interesting and relevant problem because of its non-trivial impact on response time (that is, the time between when a request arrives and when it has been handled).

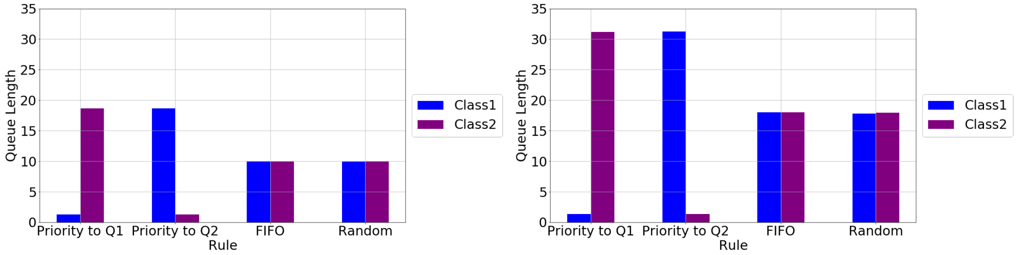
This tutorial paper considers scheduling in the context of a multiclass many-server queue. In this queue, customers request service and servers are the resources available. Each server can provide service to at most one customer at a time. This is a one-pass system, meaning that each customer sees a server at most once. Waiting customers are impatient, and will abandon the system without receiving service if their wait becomes too long.

* Research supported in part by NSF Grant DMS-1510198.

[†] Research supported in part by the William S. Fishman Faculty Research Fund at the University of Chicago Booth School of Business.

The notation commonly used to describe the multiclass many-server queue we study is $G/GI/N+GI$. The first "G" refers to the class specific customer arrival processes that are general counting processes in the sense that they can be time-varying and only need satisfy mild restrictions (see (4) and the surrounding paragraph). The second "GI" refers to the fact that the class specific amounts of time required for servers to process customer requests are independent and can be described by any general distribution that satisfies mild restrictions (see Assumption 1). The third "GI" refers to the class-specific patience distributions that govern how long customers will wait to enter service before abandoning, which are independent and satisfy mild restrictions (see Assumption 1). The independence assumption is more natural for situations in which the customers cannot observe the queue than for situations in which they can. The arrival processes, service times, and patience times are all independent of each other[1]. The "$N \in \{1,2,\ldots\}$" is the number of servers, which implies that up to N customers can be processed in parallel. If $N+n$ customers are present in the system for any positive integer n, then at least n customers must be waiting (and more than n customers will be waiting if servers can idle in the presence of waiting customers).

The scheduling policy determines the sequence in which waiting customers are served. If customers are homogeneous, then a common scheduling policy is first-come-first-served (FCFS), which serves customers in the order in which they arrive. However, in general, customers are heterogeneous, and have different arrival patterns, different processing requirements, and different waiting time behaviors, as well as being of different importance. We would like to have methodology that gives rise to scheduling policies that account for customer heterogeneity.



(a)Exponential(1) patience distribution.

(b)Uniform(0,2) patience distribution.

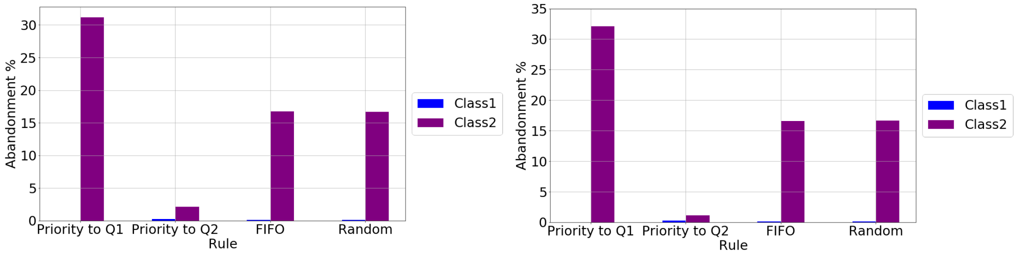
Figure 1. The simulated average queue size in the $G/GI/N+GI$ queue with two classes, each having Poisson arrivals with rate 60 arrivals per time unit, $N=100$ servers, and Exponential service time distributions with mean equal to one time unit.

Figure 1 emphasizes that the scheduling policy is a first order determinant of the quality of service in a system in which the customers are grouped into two different classes[2] . (Exponential(1) refers to an exponential distribution with mean 1 time unit and Uniform(0,2) refers to a uniform distribution with lower bound 0 and upper bound 2 time units.) When priority is given to one class over the other, almost all of the adverse effects associated with congestion (such as waiting) are experienced by the low priority class. "Fair" policies, such as FCFS and randomizing over which class receives service by flipping an unbiased coin when

[1] Although this is common in the queueing literature to assume such independence, correlation between service and patience times has been observed empirically [55], and is studied in the recent paper [65].

[2] Each simulation shown in Figure 1, and later in Figure 2, is run until 5 million customers arrive, with the time before the arrival of the 1000th customer considered a "warm-up" period and discarded. The numbers graphed are for one simulation run.

customers from both classes are waiting, result in both classes experiencing similar amounts of congestion. These observations are true regardless of the patience distributions. However, as can be seen by comparing Figures 1 (a) and (b), the total average queue length (that is, the number of customers waiting) is affected by the patience distribution. In contrast, Figures 2 (a) and (b) show that the percentage of customer abandonments is not affected by the patience distribution.



(a)Exponential(1) patience distribution. (b)Uniform(0,2) patience distribution.

Figure 2. The simulated average abandonment percentage in the $G/GI/N+GI$ queue with two classes, each having Poisson arrivals with rate 60 arrivals per time unit, $N=100$ servers, and Exponential service time distributions with mean equal to one time unit.

Figures 1 and 2 lead us to wonder how much the patience distribution should affect the scheduling. Figure 2 suggests that when abandonments are costly, the system manager should give priority to the more costly class, regardless of the patience distribution. However, when the system manager also cares about congestion effects such as average queue lengths (which also by Little's law determine average wait times), then Figure 1 suggests that the patience distribution may play a role. To study these questions, we use fluid approximation methodology, beginning with the case that only abandonments are penalized, and later extending to the case where queue lengths are penalized through holding costs. The value of this methodology is supported by the following rate calculation, that is consistent with fluid scale: When priority is given to one class, the other class's abandonment percentage should be approximately

$$\frac{\text{low priority class arrival rate - remaining service capacity for that class}}{\text{low priority class arrival rate}} = \frac{60-40}{60} = 33.33\%.$$

The predicted 33.33% abandonment matches the simulated percentages in Figure 2.

The main purpose of this paper is to discuss a fluid model for the multiclass many-server $G/GI/N+GI$ queue that allows the arrival process to have a time-varying rate vector and is valid for a wide range of natural scheduling policies. The fluid model is a first order, law-of-large numbers, deterministic description of the system evolution. The fluid model presented here arises in [59] as a means for characterizing fluid limit points for a general class of scheduling policies. It has origins in the single class papers [42, 44], and in the multiclass paper [9], which studies static priority scheduling. The fluid model presented here is as in [9], but with more fundamental relationships replacing the static priority policy specific equations.

The fluid model, in contrast to the discrete-event queue, is powerful because of its analytic tractability, which makes it well suited to help inform scheduling decisions. We show how to make these decisions when the fluid arrival process has constant arrival rate vector, but there are also practical benefits to presenting the more general fluid model (with time-varying rate) because of the ubiquity of time-varying arrival rates in real life (and we hope that the more general fluid model will be helpful for future research).

We characterize the fluid model invariant states when the fluid model (absent policy specific scheduling equations) has constant arrival rate vector. The fluid model invariant states are those states for which a solution to the fluid model equations remains constant in time. The importance of characterizing the invariant states is that they provide an approximation to the mean steady-state system behavior under different scheduling policies. Then, the fluid model invariant states can be used to formulate a fluid control problem that approximates a scheduling control problem of interest for the multiclass many-server queue. Looking ahead, we propose to use the fluid control problem in (23) to provide insights into what can be a good scheduling policy for the scheduling control problem stated in (1).

The remainder of this paper is organized as follows. Section 2 details a scheduling problem formulation that penalizes only abandonments. We write the fluid model equations in Section 3, and we characterize the fluid model invariant states when the arrival rate vector is constant in time in Section 4. Section 5 sets up the fluid control problem and Section 6 provides the solution. We suggest some open problems in Section 7, including a discussion of the case when congestion is penalized through holding costs. Sections 8 and 9 provide the proofs of the results stated in this paper (Theorem 1 and Lemma 1).

2. Problem Formulation

We study the multiclass $G/GI/N+GI$ queue shown in Figure 3. Customers from class $j\in\mathbb{J}:=\{1,\ldots,J\}$ arrive according to a counting process $E_j(t)$, that is independent of all other customer arrival processes. Each customer arrives with a patience time and abandons the system without being served if service is not commenced before the patience time expires. The patience time is the maximum amount of time a customer will wait in the system to begin service, and is also known as the reneging time in the literature. The N servers are fully flexible in the sense that every server can serve every customer; however, the service time may depend on the customer class. Upon arrival each class j customer independently samples from the distribution determined by cdf G^r_j having mean $1/\theta_j\in(0,\infty)$ to find his patience time and from the distribution determined by cdf G^s_j having mean $1/\mu_j\in(0,\infty)$ to find his service time, $j\in\mathbb{J}$. The superscript r is mnemonic for reneging and the superscript s is mnemonic for service.

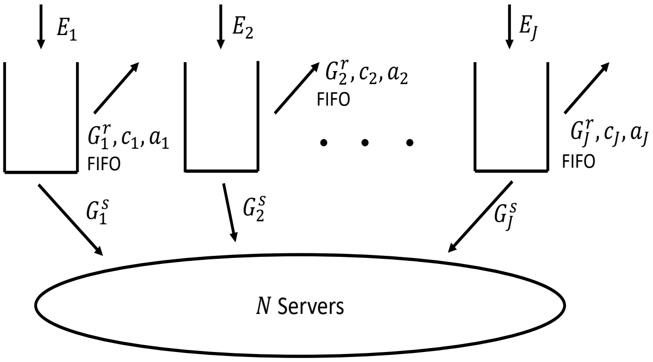


Figure 3. The V-model: Multiple customer types and fully flexible servers.

Assumption 1. *We assume E_j is a counting process (that is, a nonnegative, nondecreasing process that assumes integer values, is right-continuous with left limits, and satisfies $E_j(0)=0$) for each $j\in\mathbb{J}$. We assume G^r_j and G^s_j are absolutely continuous with density functions g^r_j and g^s_j that have (possibly infinite) right edges of support*

$$H^s_j:=\sup\{x\in[0,\infty):1-G^s_j(x)>0\}\ \textit{and}\ H^r_j:=\sup\{x\in[0,\infty):1-G^r_j(x)>0\},$$

for each $j \in \mathbb{J}$. We let

$$h_j^r(t) := \frac{g_j^r(t)}{1 - G_j^r(t)} \text{ for } t \in [0, H_j^r) \text{ and } h_j^s(t) := \frac{g_j^s(t)}{1 - G_j^s(t)} \text{ for } t \in [0, H_j^s)$$

be the associated hazard rate functions, for each $j \in \mathbb{J}$.

The scheduling policy determines what to do when a server is available. In particular, when customers from multiple classes are waiting, the scheduling policy must specify which class, if any, the server should next serve. Within the class selected, the customer who has waited the longest will be served; this customer is called the head-of-line (HL) customer. Equivalently, within any given class, the scheduling policy respects FCFS ordering. If a server becomes available to find no waiting customers, then the server necessarily idles.

Customer abandonments are a clear indication of customer dissatisfaction. Consequently, abandonments are costly. However, all abandonments may not be equally costly. For example, in a revenue-generating system, the abandonment of a customer from a higher-revenue class costs more than that of a customer from a lower-revenue class. To capture this distinction, we assume there is a class-dependent abandonment cost $a_j \in (0,\infty), j \in \mathbb{J}$. (Later, in Section 7, we will also allow a class-dependent holding cost $c_j \in (0,\infty), j \in \mathbb{J}$.) The process $R_j(T,\pi)$ tracks the cumulative number of abandonments from class $j \in \mathbb{J}$ under scheduling policy π.

Our objective is to find a scheduling policy π that minimizes the long run average cost:

$$\mathcal{C}(\pi) := \limsup_{T\to\infty} \frac{1}{T} I\!\!E\left[\sum_{j=1}^{J} a_j R_j(T,\pi)\right].$$

The class of HL scheduling policies Π that we consider are those that (i) do not assume knowledge of the future, (ii) enforce that once a customer enters service that customer stays in service until completion, and (iii) satisfy a mild condition that ensures the oscillations of the processes tracking the number of customers from each class that enter service are not too large. The class Π allows idling scheduling policies; that is, the scheduling policy can be such that servers sometimes idle when customers are waiting in the queue. (A precise mathematical statement specifying Π can be found in the first Definition in Section 2.5 in [59].) Thus, we would like to find $\pi^\star \in \Pi$ such that

$$\mathcal{C}(\pi^\star) := \inf_{\pi\in\Pi} \mathcal{C}(\pi). \tag{1}$$

The problem (1) is not amenable to exact analysis, and so we investigate approximate solutions. In particular, we use the fluid model presented in Section 3 to construct an analytically tractable approximating control problem. The fluid model is more easily understood after clearly defining the $G/GI/N+GI$ queue state space, which we do next.

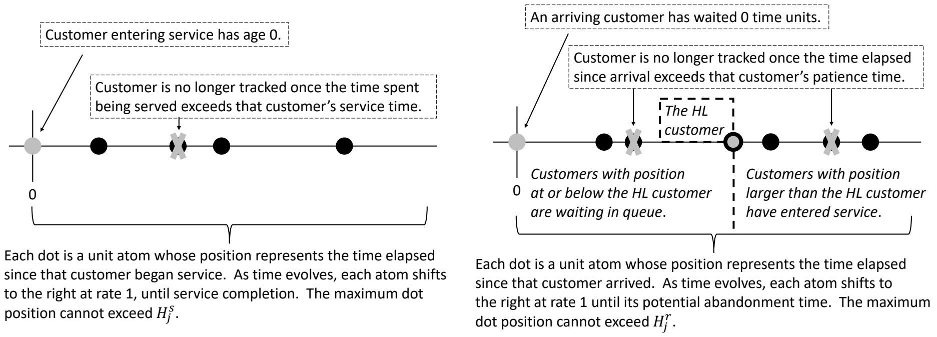
The State Space

We let $\mathbb{R} := (-\infty,\infty)$ denote the set of real numbers, $\mathbb{R}_+ := [0,\infty)$ the set of nonnegative real numbers, and $\mathbb{Z}_+ := \{0,1,2,\ldots\}$ the set of nonnegative integers. For $H \in \mathbb{R}_+ \cup \{\infty\}$, $\mathbf{M}[0,H)$ is the set of finite nonnegative Borel measures on $[0,H)$, endowed with the topology of weak convergence, which is a Polish space. The $G/GI/N+GI$ queue state at time $t \in \mathbb{R}_+$ is described as follows: For each $j \in \mathbb{J}$,

- $\alpha_j(t) \in \mathbb{R}_+$ is the time elapsed since the last class j customer arrived to the system;
- $X_j(t) \in \mathbb{Z}_+$ is the number of class j customers in the system (either waiting in the queue or being served);

- $\nu_j(t) \in \mathbf{M}[0, H_j^s)$, shown in Figure 4(a), encodes the length of time every class j customer in service at time t has been in service, and is also known as the time t age-in-service measure;
- $\eta_j(t) \in \mathbf{M}[0, H_j^r)$, shown in Figure 4(b), stores the amount of time that has passed between each class j customer's arrival time up until that customer's potential abandonment time (which is the arrival time plus the sampled patience time), for every class j customer that arrived before time t, and without regard for whether or not that customer has entered service, and is also known as the time t potential queue measure.

The measures ν_j and η_j track the evolution of unit atoms over time, where each atom is associated with a particular class j customer's time-in-service, or time-since-arrival, as shown in Figure 4.


Customer entering service has age 0.
Customer is no longer tracked once the time spent being served exceeds that customer's service time.
0
Each dot is a unit atom whose position represents the time elapsed since that customer began service. As time evolves, each atom shifts to the right at rate 1, until service completion. The maximum dot position cannot exceed H_j^s.
An arriving customer has waited 0 time units.
Customer is no longer tracked once the time elapsed since arrival exceeds that customer's patience time.
The HL customer
0
Customers with position at or below the HL customer are waiting in queue.
Customers with position larger than the HL customer have entered service.
Each dot is a unit atom whose position represents the time elapsed since that customer arrived. As time evolves, each atom shifts to the right at rate 1 until its potential abandonment time. The maximum dot position cannot exceed H_j^r.

(a)Age-in-service measure-valued process ν_j.
(b)Potential queue measure-valued process η_j.

FIGURE 4. A graphic representation of the state space measures for a given class $j \in \mathbb{J}$.

The fourth bullet point requires additional explanation. The measure η_j, $j \in \mathbb{J}$, tracks class j customers that are "potentially" waiting in the queue. Customers "potentially" in the queue are those that have arrived, but whose potential abandonment time has not passed. The term potential refers to the fact that such customers may or may not have entered and/or finished service. In particular, the potential customers waiting in queue are the HL customer and those that have waited less than the HL customer. All potential customers that have waited longer than the HL customer have entered service (and may or may not have finished service). The fact that the potential queue measure $\eta_j, j \in \mathbb{J}$, is independent of the scheduling policy is helpful for analytic tractability.

As our goal is to study an associated fluid model, we do not provide the full system dynamics required to determine how the $G/GI/N+GI$ state evolves over time here. For further detail, we refer the reader to Section 2 of [59]. Instead, we provide some idea of how other processes of interest can be derived from the state. Suppose the state at time $t \geq 0$ is $(\alpha(t), X(t), \nu(t), \eta(t))$. Then, the number of class $j \in \mathbb{J}$ customers $B_j(t)$ in service at time $t \geq 0$ is found by integrating the measure $\nu_j(t)$ over $x \in \mathbb{R}_+$ to count the number of unit atoms contained in the measure (i.e., count the number of dots on the x-axis in Figure 4 (a)), so that

$$B_j(t) = \int_0^{H_j^s} \nu_j(t)(dx). \tag{2}$$

From the above display, the number of class $j \in \mathbb{J}$ customers $Q_j(t)$ waiting in queue at time $t \geq 0$ is

$$Q_j(t) = X_j(t) - \int_0^{H_j^s} \nu_j(t)(dx) = X_j(t) - B_j(t).$$

Similar reasoning as above shows that the number of class j customers whose potential abandonment times have not yet passed at time $t \geq 0$ is

$$\int_0^{H_j^r} \eta_j(t)(dx).$$

The above display is also interpreted as the number of class j customers potentially in the queue at time $t \geq 0$, and so is an upper bound on the number of class j customers actually waiting in the queue at time $t \geq 0$; i.e.,

$$Q_j(t) = X_j(t) - B_j(t) \leq \int_0^{H_j^r} \eta_j(t)(dx). \tag{3}$$

3. The Fluid Model with Time-Varying Input

The input to the fluid model is an arrival function E having domain $\mathbb{R}_+$ and range $\mathbb{R}_+^J$. The arrival function is continuous, and each component has initial value zero and is non-decreasing. The arrival function arises as the fluid limit for a sequence of multiclass $G/GI/N+GI$ queues, as introduced in Section 2, as the number of servers tends to infinity $(N \to \infty)$ while simultaneously increasing the volume of arrivals to be of the same order, order N. More specifically, if we consider a sequence of queues, indexed by the number of servers N, and let $E_j^N, j \in \mathbb{J}$, be the class j arrival process to the queue with N servers, then the arrival function E arises from the functional strong law assumption

$$\mathbb{P}\left(\lim_{N\to\infty} \max_{j\in\mathbb{J}} \sup_{0\leq t\leq T} \left| \frac{E_j^N(t)}{N} - E_j(t) \right| = 0 \right) = 1, \text{ for any } T \in (0,\infty). \tag{4}$$

For example, if, for each $j \in \mathbb{J}$, E_j^N is a renewal process with rate $\lambda_j N$ for specified $\lambda_j \in (0,\infty)$, so that the arrival rate to each class increases linearly as the number of servers N increases, then (4) holds with $E_j(t) = \lambda_j t$, $t \in \mathbb{R}_+$, for each $j \in \mathbb{J}$. (For introductory background on the functional strong law for renewal processes, see Sections 5.4 and the beginning of Section 5.5 in [18].) As another example, if, for each $j \in \mathbb{J}$, E_j^N is a non-stationary Poisson process with time-dependent instantaneous arrival rate function $N\lambda_j(t)$ that is nonnegative and integrable on $[0,t]$ for all $t \in \mathbb{R}_+$, then (4) holds with $E_j(t) = \int_0^t \lambda_j(s)ds < \infty$ for each $j \in \mathbb{J}$ and all $t \in \mathbb{R}_+$, so that the resulting fluid arrival function has time-varying rate.

The condition (4) provides an explicit connection between the $G/GI/N+GI$ queue described in Section 2 and the fluid model we specify in this Section. Under mild asymptotic conditions, that fluid model has solutions that arise as limit points of sequences of functional law of large numbers scaled state descriptors for multiclass $G/GI/N+GI$ queues operating under any scheduling policy in the class Π; see Theorem 4.1 in [59].

The astute reader will have noticed that in the paragraph surrounding (4) we re-used the notation $E_j, j \in \mathbb{J}$. In Section 2, E_j denotes the class j arrival process to the $G/GI/N+GI$ queue. In the paragraph surrounding (4), E_j^N denoted the class j arrival process to the $G/GI/N+GI$ queue, and E_j denotes the jth component of the arrival function E for the fluid model. From this point forward, whenever we refer to a process, measure, or quantity associated with the $G/GI/N+GI$ queue, we use the superscript N. Without the superscript N, the reader should interpret the process, measure, or quantity as being associated with the fluid model.

The scaling assumed in (4) (that is, the division by N) is helpful to keep in mind to understand the intuition behind the fluid model. The result of the scaling is that the customers are no longer thought of as individual units arriving at discrete points in time but are instead thought of as a fluid that flows continuously into the system over time, and may stay in the system for a positive amount of time before departing, either through service completion or abandonment. In the remainder of this section, the term "scaled" applied to a process, measure, or quantity associated with the multiclass $G/GI/N+GI$ queue means that process is divided by N.

The fluid model state space follows from the scaled state space for the $G/GI/N+GI$ queue when the number of servers N increase to infinity and (4) holds. Because the arrival function E is continuous, we do not need to track the time elapsed since the last arrival. Due to (4), this is zero in the limit. However, the other elements of the $G/GI/N+GI$ queue state space are relevant. The fluid model state at time $t \in \mathbb{R}_+$ is described as follows: For each $j \in \mathbb{J}$,

- $X_j(t) \in \mathbb{R}_+$ approximates the scaled number of customers in the system at time t;
- $\nu_j(t) \in \mathbf{M}[0,H_j^s)$ is a measure-valued function that approximates the scaled age-in-service measure at time t;
- $\eta_j(t) \in \mathbf{M}[0,H_j^r)$ is a measure-valued function that approximates the scaled potential queue measure at time t.

We endow the product space $\mathbb{R}_+^J$ with the usual Euclidean topology. We set

$$\mathbb{X} := \mathbb{R}_+^J \times \left(\times_{j=1}^J \mathbf{M}[0,H_j^s)\right) \times \left(\times_{j=1}^J \mathbf{M}[0,H_j^r)\right),$$

and endow $\mathbb{X}$ with the product topology, which is a Polish space (recalling that $\mathbf{M}[0,H_j^s)$ and $\mathbf{M}[0,H_j^r)$, for $j \in \mathbb{J}$, are endowed with the topology of weak convergence).

A fluid model solution satisfies the evolution equations presented below, as well as extra conditions, and is defined precisely following those evolution equations in Definition 1. Any fluid model solution is in $\mathbf{C}(\mathbb{X})$, which denotes the set of functions having domain $\mathbb{R}_+$ and range $\mathbb{X}$ that are continuous in time. Any fluid model solution is in $\mathbf{C}(\mathbb{X})$.

We require the following notation to present a fluid model solution. For any Borel measurable function f, having domain $[0,H)$ for specified $H \in (0,\infty]$ and range $\mathbb{R}$, that is either nonnegative or integrable with respect to the measure $\xi \in \mathbf{M}[0,H)$, let

$$\langle f, \xi \rangle := \int_0^H f(x)\xi(dx).$$

For a simple example, if $f(x) = x$ for all $x \in \mathbb{R}_+$ is the identity function, and ξ is the exponential distribution with rate parameter θ, then

$$\langle f, \xi \rangle = \int_0^\infty x\theta \exp(-\theta x)dx = \frac{1}{\theta},$$

which is its mean.

Given an arrival function E, a fluid model solution $(X,\nu,\eta) \in \mathbf{C}(\mathbb{X})$ for E satisfies the conditions

$$\int_0^t \left\langle h_j^s, \nu_j(u) \right\rangle du < \infty \qquad \text{and} \qquad \int_0^t \left\langle h_j^r, \eta_j(u) \right\rangle du < \infty, \tag{5}$$

which ensure the cumulative amount of fluid that has abandoned and departed is finite for all $t \in \mathbb{R}_+$ (see (10) and (11) below), has an initial potential queue measure with no atoms

$$\left\langle 1_{\{x\}}, \eta_j(0) \right\rangle = 0 \text{ for all } x \in [0,H_j^r), \tag{6}$$

and has auxiliary functions (defined mathematically below) interpreted as follows: For each $j \in \mathbb{J}$, at time $t \in \mathbb{R}_+$,

- $B_j(t)$ approximates the scaled number of customers from each class in service;
- $Q_j(t)$ approximates the scaled number of customers from each class waiting in queue;

- $R_j(t)$ approximates the scaled cumulative number of customers from each class that have abandoned in $[0,t]$;
- $D_j(t)$ approximates the scaled cumulative number of customers from each class that have departed after being served in $[0,t]$;
- $K_j(t)$ approximates the scaled cumulative number of customers from each class that have entered service in $[0,t]$, and depends on scheduling policy.

The notation $\mathbf{C}(\mathbb{R}_+^J)$ denotes the set of functions having domain $\mathbb{R}_+$ and range $\mathbb{R}_+^J$ that are continuous, endowed with the usual Skorokhod J_1-topology [13]. The auxiliary functions B, Q, R, D, K are all in $\mathbf{C}(\mathbb{R}_+^J)$.

The measure $\nu(t)$ determines the amount of fluid in service at time t from each class, and so

$$B_j(t) := \langle 1, \nu_j(t) \rangle = \int_0^{H_j^s} \nu_j(t)(dx), \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+ \tag{7}$$

The expression in (7) is as in (2). The difference in interpretation is that the division by N has resulted in a total service capacity of 1 (instead of N), and $B_j(t)$ is interpreted as the fraction of service capacity devoted to customer class j at time t.

Fluid present in the system must either be in service or waiting, and so

$$Q_j(t) := X_j(t) - B_j(t), \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+, \tag{8}$$

which we require to be nonnegative for all $j \in \mathbb{J}$ and all $t \in \mathbb{R}_+$. The amount of fluid waiting is bounded above by the amount of fluid potentially in queue; that is, consistent with the upper bound inequality in (3), we require

$$Q_j(t) \leq \langle 1, \eta_j(t) \rangle, \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+.$$

We use the measure $\eta_j(t)$ to derive the cumulative amount of class j fluid that abandons by time t without ever having entered service. To see how to do this, first recall from the explanation of the potential queue measure, that the measure-valued function η_j tracks two types of fluid, fluid waiting in queue and fluid that has entered service, but whose potential abandonment time has not passed. Next, recognize that the assumption of a HL scheduling policy ensures fluid is ordered by its age. Then, the function χ_j given by

$$\chi_j(t) := \inf\{y \in \mathbb{R}_+ : \langle 1_{[0,y]}, \eta_j(t) \rangle \geq Q_j(t)\}, \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+, \tag{9}$$

represents the waiting time of the oldest class j fluid in queue at time t, and all fluid having age less than $\chi_j(t)$ is waiting in queue while that having age greater than $\chi_j(t)$ is no longer waiting in queue (because that fluid has entered service). We note that due to the continuity of E and (6), $\eta_j(t)$ has no atoms for all $t \geq 0$ and $j \in \mathbb{J}$, which implies

$$\langle 1_{[0,\chi_j(t)]}, \eta_j(t) \rangle = Q_j(t) \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+;$$

see Remark 1 below. For any $j \in \mathbb{J}, t \in \mathbb{R}_+$, and $w \in [0, \chi_j(t)]$, the value of the hazard rate $h_j^r(w)$ dictates the rate at which fluid with age w abandons in the next instant. This leads to cumulative abandonment function

$$R_j(t) := \int_0^t \left(\int_0^{\chi_j(u)} h_j^r(w) \eta_j(u)(dw) \right) du, \text{ for } j \in \mathbb{J} \text{ and all } t \in \mathbb{R}_+. \tag{10}$$

For the concrete example when the patience distribution is exponential with mean $1/\theta_j$ and the system is initially empty, (10) reduces to the simpler expression

$$R_j(t) = \int_0^t \theta_j \left(\int_0^{\chi_j(u)} \eta_j(u)(dw) \right) du = \int_0^t \theta_j Q_j(u) du, \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+.$$

Class j fluid that has been in service $x \in \mathbb{R}_+$ time units departs in the next instant at rate $h_j^s(x)$, which leads to

$$\langle h_j^s, \nu_j(u) \rangle = \int_0^{H_j^s} h_j^s(x) \nu_j(u)(dx)$$

being the class j instantaneous departure rate at time $u \in \mathbb{R}_+$. Then, the cumulative amount of class j fluid that has departed by time t after completing service is

$$D_j(t) := \int_0^t \langle h_j^s, \nu_j(u) \rangle \, du, \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+. \tag{11}$$

Mass conservation implies a natural balance relationship between the entry-into-service function, the cumulative departure function, and the amount of fluid in service function. This explains the definition

$$K_j(t) := B_j(t) + D_j(t) - B_j(0), \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+. \tag{12}$$

Finally, in order to mathematically define a fluid model solution, we must specify constraints on the evolution of $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ and its associated auxiliary functions $B, Q, R, D, K \in \mathbf{C}(\mathbb{R}_+^J)$. First, the amount of fluid in the system must respect the conservation of flow equation

$$X_j(t) = X_j(0) + E_j(t) - R_j(t) - D_j(t), \text{ for all } j \in \mathbb{J} \text{ and } t \in \mathbb{R}_+. \tag{13}$$

Second, for any continuous and bounded function f having domain $\mathbb{R}_+$, the measure-valued functions ν and η evolve over time according to the following equations: For all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$,

$$\langle f, \nu_j(t) \rangle = \int_0^{H_j^s} f(x+t) \frac{1 - G^s(x+t)}{1 - G^s(x)} \nu_j(0)(dx) + \int_0^t f(t-u)(1 - G_j^s(t-u)) dK_j(u), \tag{14}$$

$$\langle f, \eta_j(t) \rangle = \int_0^{H_j^r} f(x+t) \frac{1 - G^r(x+t)}{1 - G^r(x)} \eta_j(0)(dx) + \int_0^t f(t-u)(1 - G_j^r(t-u)) dE_j(u). \tag{15}$$

In (14), the first term tracks fluid departing from the system that was in service at time zero, while the second term tracks when fluid that entered service after time zero departs the system. In (15), the first term tracks when the patience time of every customer that is initially present in the system at time zero expires, while the second tracks that of arriving fluid, without regard for service entry.

Definition 1. Given an arrival function E, a fluid model solution for E is $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ that satisfies (5) and (6), has auxiliary functions defined by (7)-(12) that satisfy $B, Q, R, D, K \in \mathbf{C}(\mathbb{R}_+^J)$, and is such that the following conditions hold:

(a) The total amount of fluid in service never exceeds capacity, $\sum_{j=1}^J B_j(t) \in [0,1]$ for all $t \in \mathbb{R}_+$;
(b) The function Q satisfies $0 \le Q_j(t) \le \langle 1, \eta_j(t) \rangle$ for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$, and the function X satisfies (13);
(c) The function K_j tracking the cumulative amount of class j fluid to have entered service in $[0,t]$ is non-decreasing for all $j \in \mathbb{J}$;
(d) The measure-valued functions ν and η satisfy (14) and (15).

A fluid model solution associated with a particular initial state is not unique. Uniqueness requires a more precise specification of the entry-into-service function K, that would connect to a defined scheduling policy in the class Π for the many-server queue. In order to use fluid model solutions to develop a control problem whose solution can provide insight into scheduling policies that perform well with respect to the scheduling problem (1), the reader can think of replacing the optimization over the scheduling policy in (1) with an optimization over a decision variable that uniquely dictates the entry-into-service function K.

Remark 1. (Discontinuous Fluid Model) We have restricted the arrival function E and the entry-into-service function K to be continuous. A more general fluid model that does not require such continuity is presented in [59]. One reason to make such a restriction in this paper is that under the condition (6), $\eta_j(t)$ has no atoms for all $j \in J$ and $t \in \mathbb{R}_+$; see [59, Lemma 3.4]. Then, for all $t \geq 0$ and $j \in \mathbb{J}$,

$$R_j(t) = \int_0^t \theta_j \left\langle 1_{[0,\chi_j(u))}, \eta_j(u) \right\rangle du = \int_0^t \theta_j \left\langle 1_{[0,\chi_j(u)]}, \eta_j(u) \right\rangle du.$$

Since no such subtlety occurs with any of the auxiliary function definitions involving ν, there is no need to require a condition equivalent to (6) for $\nu(0)$.

Remark 2. (Uniqueness when $J=1$) In the case of a single customer class ($J=1$), the solution to the scheduling problem is trivial. The optimal policy is to not allow servers to idle when customers are waiting. This amounts to adding the standard non-idling equation $\max(X_1(t)-1,0) = Q_1(t)$ for all $t \in \mathbb{R}_+$. Then, [42, Theorem 3.5] proves that for reasonable initial states there exists a unique solution to the fluid model equations, and [42, Theorem 3.6] establishes weak convergence to that solution for the scaled $G/GI/N+GI$ state processes under mild asymptotic conditions; see also [68, Theorems 3.5 and 4.4] for an extension that does not require the absolute continuity of the service and patience distributions, [67, Theorems 3.1 and 3.3] for an alternative approach (involving residual times rather than ages) to developing similar results, and [41] for the equivalence of the two approaches. In contrast, the non-uniqueness of a fluid model solution for the multiclass case is necessary to be able to use the fluid model to formulate a control problem that approximates the scheduling problem (1).

Remark 3. (Static Priority Scheduling) The fluid model in [9] is relevant for a multiclass $G/GI/N+GI$ queue that operates under a static priority scheduling rule. That fluid model is consistent with what is presented here, except additional equations that restrict the entry-into-service function K to that arising under static priority are added. Then, the existence of a fluid model solution follows from the existence of a solution to the fluid model in [9]; see [9, Remark 3.1(a) and Theorem 4.3]. Furthermore, the fluid model in [9] with specified initial state has a unique solution; see [9, Theorem 3.1]. In comparison, in order to present a fluid model relevant for a wider range of scheduling rules, we provide minimal restrictions on the entry-into-service function K, meaning there is no expectation of uniqueness.

Remark 4. (HL Scheduling) The assumption of HL scheduling is common, but further thought is warranted. In the single class ($J=1$) setting, HL scheduling (equivalently, FCFS) minimizes the fluid queue-lengths [12, Corollary 1] when the patience distribution has decreasing hazard rate. In contrast, the scheduling policy that prioritizes the customer that has waited the least amount of time (that is, last-come-first-serve) minimizes the fluid queue-lengths when the hazard rate is increasing [12, Corollary 1]. When the hazard rate is quasi-concave, the optimal scheduling is either HL or last-come-first-serve, depending on whether or not the hazard rate is such that a specified condition is satisfied [12, Proposition 5].

Remark 5. (A No Abandonment Model) For a multiclass $G/GI/N$ queue with no abandonments, the relevant fluid model eliminates the measure η, and has state space $\tilde{\mathbb{X}} := \mathbb{R}_+^J \times \left(\times_{j=1}^J \mathbf{M}[0,H_j^s)\right)$. Then, Definition 1 is modified as follows: Given an arrival function E, a fluid model solution for E is $(X,\nu) \in \mathbf{C}(\tilde{\mathbb{X}})$ that satisfies $\int_0^t \left\langle h_j^s, \nu_j(u) \right\rangle du < \infty$, has auxiliary functions defined by (7), (8), (11), and (12) that satisfy $B, D, Q, K \in \mathbf{C}(\mathbb{R}_+^J)$, and is such that (14), (15), the conditions (a) and (c) in Definition 1 hold, and (13) is satisfied with $R_j(t) = 0$ for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$. When $J=1$ and with the standard non-idling equation given in Remark 2 added, this is consistent with the fluid model in [44].

4. Fluid Model Invariant States

The scheduling problem (1) minimizes the long run average cost. The long run average cost is determined by the steady-state abandonment rate, assuming the existence of a unique system steady-state and attractiveness to that steady-state as time becomes large. This requires that each class's arrival process has constant rate, which occurs when we assume that the inter-arrival times to the $G/GI/N+GI$ queue are independently sampled from a given distribution. Then, the arrival processes are renewal, and the queue is a multiclass $GI/GI/N+GI$ queue. Recall that in this case the condition (4) holds when the renewal arrival processes have rate $\lambda_j N$ for each $j \in \mathbb{J}$. The associated fluid model has linear arrival function E; that is, each component of E is defined from a specified constant $\lambda_j \in \mathbb{R}_+$ as

$$E_j(t) := \lambda_j t \text{ for all } t \geq 0, \text{ and each } j \in \mathbb{J}. \tag{16}$$

For the remainder of this paper, we assume (16) holds, unless explicitly stated otherwise. We call $\lambda := (\lambda_1, \ldots, \lambda_J)$ the arrival rate vector.

In the single class case, [43, Theorem 3.3] establishes that the system steady-state, when appropriately scaled, converges to the unique fluid model invariant state, under mild assumptions[3]. This suggests that the multiclass $GI/GI/N+GI$ queue steady-state, when appropriately scaled, is well approximated by the invariant states of the fluid model given in Section 3. This motivates us to characterize the invariant states of the fluid model, from which we formulate a fluid control problem to approximate (1).

Definition 2. A tuple $(X^*, \nu^*, \eta^*) \in \mathbb{X}$ is said to be an invariant state for the arrival function (16) if the constant function (X, ν, η) given by $(X(t), \nu(t), \eta(t)) = (X^*, \nu^*, \eta^*)$ for all $t \geq 0$ is a fluid model solution (i.e., satisfies Definition 1). We let $\mathcal{I}_\lambda$ denote the set of invariant states for λ. For any invariant state $(X^*, \nu^*, \eta^*) \in \mathcal{I}_\lambda$, we let $(B^*, Q^*, R^*, D^*, K^*)$ denote the associated auxiliary functions.

The characterization of the invariant states uses the inverse of the patience distribution function and its associated excess life distribution. The excess life distribution characterizes the amount of time remaining until the next event for a renewal process in stationarity (see, for example, Example 7.24 in [61]) and is given by

$$G^r_{e,j}(x) := \int_0^x \theta_j (1 - G^r_j(y)) dy, \text{ for } j \in \mathbb{J} \text{ and } x \in \mathbb{R}_+,$$

recalling that $1/\theta_j$ is the mean of the patience distribution.

An invariant state is uniquely specified by the allocation of server capacity given to each class. Those allocations must lie in the set

$$\mathbb{B} := \left\{ b \in \mathbb{R}^J_+ : b_j \leq \rho_j \text{ for all } \ j \in \mathbb{J} \text{ and } \sum_{j=1}^J b_j \leq 1 \right\},$$

where, for each $j \in \mathbb{J}$,

$$\rho_j := \lambda_j / \mu_j$$

is the instantaneous fluid workload contribution from class j. The set $\mathbb{B}$ captures the long run average fraction of the collective server effort that could be provided to each class. A suitable specification of the entry-into-service function K results in a unique $b \in \mathbb{B}$; however, there may be many entry-into-service functions that give rise to the same b. Under reasonable asymptotic conditions, the limit as $N \to \infty$ of the rescaled cost $C^N(\pi^N)/N$ should give rise to a unique $b \in \mathbb{B}$ that achieves the limiting cost, and that b uniquely specifies the fluid model invariant states from Theorem 1 below.

[3] Note that a forthcoming proof correction may require some additional assumptions not present in the original manuscript.

Theorem 1. *(Fluid Model Invariant States) Suppose that G_j^r is strictly increasing for each $j \in \mathbb{J}$, with inverse function $\left(G_j^r\right)^{-1}$. For $b \in \mathbb{B}$, define*

$$q_j(b_j) := \begin{cases} \frac{\lambda_j}{\theta_j}, & \text{if } b_j = 0, \\ \frac{\lambda_j}{\theta_j} G_{e,j}^r\left(\left(G_j^r\right)^{-1}\left(1 - \frac{b_j}{\rho_j}\right)\right), & \text{if } b_j \in (0, \rho_j] \end{cases} \quad \text{for } j \in \mathbb{J}, \tag{17}$$

and, for each $j \in \mathbb{J}$, let

(i) $\eta_j^*(dx) := \lambda_j \left(1 - G_j^r(x)\right) dx$ *for each* $x \in \mathbb{R}_+$,
(ii) $\nu_j^*(dx) := b_j \mu_j \left(1 - G_j^s(x)\right) dx$ *for each* $x \in \mathbb{R}_+$,
(iii) $X_j^* := b_j + q_j(b_j)$.

Then $(X^*, \nu^*, \eta^*) \in \mathcal{I}_\lambda$, $B^* = b$, and $Q^* = q$. Conversely, if $(X^*, \nu^*, \eta^*) \in \mathcal{I}_\lambda$, then $B^* \in \mathbb{B}$ and (X^*, ν^*, η^*) satisfies (i)-(iii) with $b = B^*$.

The proof of Theorem 1 is found in Section 8. In the case of one customer class ($J = 1$) and service rate $\mu = 1$, the expression in Theorem 1 for q_1 is consistent with [64, Equation (3.7)].

Henceforth, we assume the conditions for Theorem 1 hold; that is, we assume G_j^r is strictly increasing for each $j \in \mathbb{J}$, with inverse function $\left(G_j^r\right)^{-1}$.

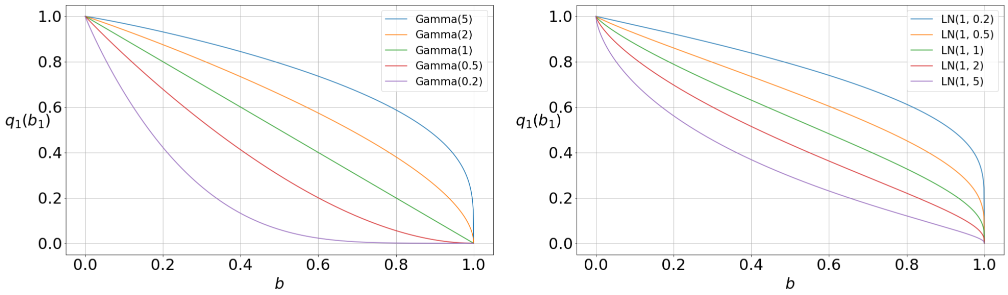
Remark 6. (Related Results) A version of Theorem 1 is proved in [43, Theorem 5.5] and in [67, Theorem 3.2] when $J = 1$ under the non-idling condition in Remark 2. A version of Theorem 1 is also proved in [9, Theorem 3.3] when the fluid model has added equations relevant for a static priority scheduling policy. In both cases, not all $b \in \mathbb{B}$ can be achieved. This motivates the need to understand the fluid limits associated with a wider range of scheduling policies.

Remark 7. (Convergence to Invariant States) The long-time behavior of the fluid model is nontrivial. In particular, we would like to know that any fluid model solution converges to an invariant state as time becomes large, under mild conditions on the initial state. When $J = 1$, this is shown in [51, Theorems 1 and 2], but requires somewhat restrictive conditions on the initial state when there is not enough capacity to serve all fluid; see also [43, Section 7.1] for more discussion of the relevant issues. Proving such a convergence result for $J > 1$ is an open problem.

Remark 8. (Prediction) For prediction purposes, recalling (4), if the arrival rate to the multiclass $G/GI/N + GI$ queue is written in terms of the number of servers N, so that $\lambda_j N$ is the class j arrival rate, then the class j fluid arrival rate is λ_j, as shown in (16). This results in the predicted class $j \in \mathbb{J}$ queue size $N q_j(b_j)$ when the scheduling policy is such that on average $b_j N$ servers are busy serving class j. Following the asymptotic regime in [59], the service rate μ_j and mean patience times $1/\theta_j$ appearing in the formula for q_j are not scaled, $j \in \mathbb{J}$.

Figure 5 shows the fluid queue sizes are ordered by the variance of the patience distributions, for $J = 1$ and $\lambda_1 = \mu_1$. Figure 5(a) assumes the patience distribution is a mean 1 Gamma distribution having both shape and rate parameters equalling $p \in \{0.2, 0.5, 1, 2, 5\}$[4]. The variance of a Gamma(p) distribution is $1/p$, which is decreasing in p. Figure 5(b) assumes the patience distribution is Lognormal(m, v), where $m = 1$ is the mean and $v \in \{0.2, 0.5, 1, 2, 5\}$ is the variance. For all distributions, the queue-length when no service effort is expended on the class ($b = 0$) equals the mean of the distribution, 1, and the queue-length when full service effort is expended on the class ($b = 1$) is zero. Otherwise, the more variable distributions result in lower queue-lengths, which has intuition as follows. The greater

[4] The cdf G for the mean 1 Gamma(p) distribution is: $G(x) = \frac{\int_0^{px} t^{p-1} e^{-t} dt}{\int_0^{\infty} t^{p-1} e^{-t} dt}$



(a)Gamma patience distributions.

(b)Lognormal patience distributions.

FIGURE 5. For $\mathbb{J}=\{1\}$, the effect of different patience distributions on $q_1(b_1)$ for $b_1 \in [0,\rho_1=1]$ when $\lambda_1=\mu_1=1$.

variability leads to the fraction of fluid that abandons having more mass concentrated on smaller patience times, meaning abandonment decisions are made more quickly after arriving, reducing congestion.

Given $b \in \mathbb{B}$, we subscript the unique invariant state and associated auxiliary functions with b when we want to emphasize that dependence. In particular, we sometimes write $(X_b^*, \nu_b^*, \eta_b^*)$ and $(B_b^*, Q_b^*, R_b^*, D_b^*, K_b^*)$.

5. The Fluid Control Problem

Theorem 1 suggests considering the scheduling problem approximation

$$\inf_{b\in\mathbb{B}} \limsup_{T\to\infty} \frac{1}{T}\sum_{j=1}^{J} a_j R_{b,j}^*(T). \tag{18}$$

When starting from the invariant state associated with $b \in \mathbb{B}$, class j fluid arrives at rate λ_j and departs via service completion at rate $b_j\mu_j$. Hence flow balance dictates that class j fluid abandons at rate $\lambda_j - b_j\mu_j$, and so we expect

$$\frac{1}{T}R_{b,j}^*(T) = \lambda_j - b_j\mu_j, \text{ for all } T\in\mathbb{R}_+, \text{ and each } j\in\mathbb{J}. \tag{19}$$

Substituting the equality in (19) into (18) leads to the analytically tractable fluid control problem

$$\min_{b\in\mathbb{B}} \sum_{j=1}^{J} a_j\left(\lambda_j - b_j\mu_j\right), \tag{20}$$

whose solution is given in the next section.

The remainder of this section shows the derivation of (19) from the invariant state given in Theorem 1. For this derivation, we do not use the b subscript in the invariant state notation. From (10) and the fact that an invariant state is constant in time,

$$R_j^*(T) = T\int_0^{\chi_j^*} h_j^r(w)\eta_j^*(dw) \text{ for } j\in\mathbb{J} \text{ and all } T\in\mathbb{R}_+. \tag{21}$$

We derive the expression for χ_j^* for any $j\in\mathbb{J}$ in this paragraph, which is necessary to simplify (21). Fix $j\in\mathbb{J}$. For this, from (9) and the fact that η_j^* has no atoms (recall Remark 1), χ_j^* solves

$$\left\langle 1_{[0,\chi_j^*]}, \eta_j^* \right\rangle = Q_j^*.$$

Equivalently, substituting the definitions of η_j^* and Q_j^* given in Theorem 1, if $b_j = 0$, then $\left\langle 1_{[0,\chi_j^*]}, \eta_j^* \right\rangle = \lambda_j/\theta_j$, which implies $\chi_j^* = H_j^r$. Otherwise, $0 < b_j \le \rho_j$ and we have

$$\int_0^{\chi_j^*} \lambda_j \left(1 - G_j^r(y)\right) dy = \frac{\lambda_j}{\theta_j} G_{e,j}^r \left(\left(G_j^r\right)^{-1} \left(1 - \frac{b_j}{\rho_j}\right)\right).$$

Substituting for the excess life distribution yields

$$G_{e,j}^r \left(\left(G_j^r\right)^{-1} \left(1 - \frac{b_j}{\rho_j}\right)\right) = \int_0^{\left(G_j^r\right)^{-1}\left(1-\frac{b_j}{\rho_j}\right)} \theta_j \left(1 - G_j^r(y)\right) dy.$$

The previous two displays imply χ_j^* satisfies

$$\int_0^{\chi_j^*} \left(1 - G_j^r(y)\right) dy = \int_0^{\left(G_j^r\right)^{-1}\left(1-\frac{b_j}{\rho_j}\right)} \left(1 - G_j^r(y)\right) dy.$$

This together with $\chi_j^* \le H_j^r$ and $\left(G_j^r\right)^{-1}(1 - b_j/\rho_j) \le H_j^r$ implies

$$\chi_j^* = \begin{cases} \left(G_j^r\right)^{-1} \left(1 - \frac{b_j}{\rho_j}\right), & \text{if } 0 < b_j \le \rho_j, \\ H_j^r, & \text{if } b_j = 0. \end{cases} \tag{22}$$

Finally, to see (19), we substitute η_j^* given in Theorem 1 into (21) to find

$$R_j^*(T) = T \int_0^{\chi_j^*} h_j^r(y) \lambda_j \left(1 - G_j^r(y)\right) dy.$$

The definition of the hazard rate, the expression for χ_j^* in (22), and straightforward calculation show that

$$\begin{aligned} R_j^*(T) &= T \int_0^{\chi_j^*} \lambda_j g_j^r(y) dy \\ &= T\lambda_j G_j^r(y)\Big|_0^{\chi_j^*} \\ &= T\lambda_j \left(1 - \frac{b_j}{\rho_j}\right). \end{aligned}$$

Recalling $\rho_j = \lambda_j/\mu_j$ implies

$$\lambda_j \left(1 - \frac{b_j}{\rho_j}\right) = \lambda_j - b_j\mu_j,$$

so that dividing by T in the above sequence of equalities yields (19).

6. Static Priority Scheduling

Recall the approximating fluid control problem in (20) and set

$$m^* := \min_{b \in \mathbb{B}} \sum_{j=1}^J a_j \left(\lambda_j - b_j\mu_j\right) \ge 0. \tag{23}$$

We denote a solution to the linear program (LP) in (23) by b^*.

We assume

$$\sum_{j=1}^J \rho_j = \sum_{j=1}^J \frac{\lambda_j}{\mu_j} > 1. \tag{24}$$

Otherwise, the solution to (23) is trivial. In particular, $b_j^* = \rho_j$ has $b^* \in \mathbb{B}$ and $m^* = 0$, so is feasible and attains the minimum possible objective function value of zero.

The LP (23) is equivalently written as

$$m^* := \sum_{j=1}^{J} a_j \lambda_j - \max_{b \in \mathbb{B}} \sum_{j=1}^{J} a_j \mu_j b_j.$$

Then, if we assume the classes are labeled so that

$$a_1\mu_1 > a_2\mu_2 > \cdots > a_J\mu_J,$$

the re-writing of the LP makes clear that its solution is to assign the maximum server effort required to ensure no long run average cost for abandonments, $b_j = \rho_j$, to as many of the classes with lower index as possible. More precisely, define

$$j^* := \min\left\{ k \in \mathbb{J} : \sum_{j=1}^{k} \rho_j > 1 \right\},$$

which satisfies $j^* \leq J$ under the assumption (24). Then,

$$b^* = \left(\rho_1, \ldots, \rho_{j^*-1}, 1 - \sum_{j=1}^{j^*-1} \rho_j, 0, \ldots, 0 \right)$$

solves (23), and the associated minimum objective function value is

$$m^* = a_{j^*}\left(\lambda_{j^*} - \mu_{j^*}\left(1 - \sum_{j=1}^{j^*-1} \rho_j \right) \right) + \sum_{j=j^*+1}^{J} a_j \lambda_j.$$

In words, the classes in the set $\{1, \ldots, j^* - 1\}$ are fully served, the class j^* is partially served, and the classes in the set $\{j^* + 1, J\}$ are not served at all, in an asymptotic sense.

The associated fluid queue-lengths can be found from Theorem 1 and are

- $q_j\left(b_j^*\right) = 0$ for $j \in \{1, \ldots, j^* - 1\}$;
- $q_{j^*}\left(b_{j^*}^*\right) = \frac{\lambda_{j^*}}{\theta_{j^*}} G_{e,j^*}^r \left(\left(G_{j^*}^r\right)^{-1} \left(1 - \frac{b_{j^*}^*}{\rho_{j^*}} \right) \right) \in \left(0, \frac{\lambda_{j^*}}{\theta_{j^*}} G_{e,j^*}^r (H_{j^*}^r)\right]$;
- $q_j\left(b_j^*\right) = \frac{\lambda_j}{\theta_j}$ for $j \in \{j^* + 1, \ldots, J\}$.

We can double-check that the solution matches the one given in (19) and (20) in [6] for an overloaded multiclass $M/M/N+M$ queue. In particular, if the patience distribution of class $j \in \mathbb{J}$ is exponential, then $q_{j^*}(b_j^*) = 0$ for $j \in \{1, \ldots, j^* - 1\}$, and

$$q_{j^*}(b_{j^*}^*) = \frac{\lambda_{j^*} - b_{j^*}^* \mu_{j^*}}{\theta_{j^*}} \text{ and } q_j\left(b_j^*\right) = \frac{\lambda_j}{\theta_j} \text{ for } j \in \{j^* + 1, \ldots, J\}. \tag{25}$$

In words, the fluid queue length for each class j equals the amount of incoming fluid not served multiplied by the mean patience time for that class.

The associated fluid queues are not in general linear, as can be seen by performing the above calculation when the patience distribution of class $j \in \mathbb{J}$ is uniform with lower bound 0 and upper bound $2/\theta_j$ (so that the mean of the class j patience distribution is $1/\theta_j$). In that case,

$$G_j^r(x) = \frac{\theta_j}{2} x, \text{ for } x \in \left[0, \frac{2}{\theta_j}\right] \text{ and } (G_j^r)^{-1}(x) = \frac{2}{\theta_j} x, \text{ for } x \in [0,1] \text{ and } j \in \mathbb{J}.$$

The associated excess life distribution is

$$G^r_{e,j}(x) = \int_0^x \theta_j \left(1 - \frac{\theta_j}{2} y\right) dy = \theta_j x \left(1 - \frac{\theta_j}{4} x\right), \text{ for } x \in \left[0, \frac{\theta_j}{2}\right] \text{ and } j \in \mathbb{J}.$$

We then calculate

$$G^r_{e,j}\left(\left(G^r_j\right)^{-1}(x)\right) = 2x - x^2, \text{ for } x \in [0,1] \text{ and } j \in \mathbb{J},$$

which shows

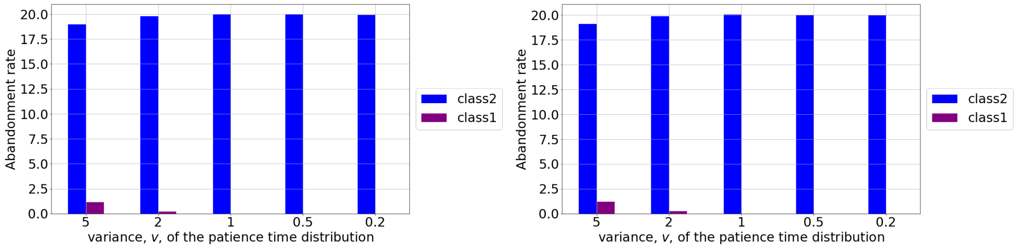
$$q_{j^*}(b^*_{j^*}) = \frac{\lambda_j}{\theta_j}\left[2\left(1 - \frac{b^*_{j^*}}{\rho_j}\right) - \left(1 - \frac{b^*_{j^*}}{\rho_j}\right)^2\right] = \frac{\lambda_j}{\theta_j}\left(1 - \left(\frac{b^*_j}{\rho_j}\right)^2\right) \tag{26}$$

and

$$q_j\left(b^*_j\right) = \frac{\lambda_j}{\theta_j} \text{ for } j \in \{j^*+1, \ldots, J\}.$$

Interpreting the Fluid Control Problem Solution as a Scheduling Policy

The fluid solution b^* that solves (23) suggests that a static priority scheduling policy π_{sp} should approximately solve the original scheduling problem (1). In particular, when a server has finished helping a customer, $\pi_{sp} \in \Pi$ assigns priority to the class having waiting customer with lowest index j (which corresponds to higher $a_j\mu_j$ values), and serves the HL customer in that class. This is exactly the well-known $c\mu$-rule, except with the value of c modified to be the class abandonment cost instead of the class holding cost.



(a)Gamma(2) service distribution.

(b)$LN(1,4)$ service distribution.

FIGURE 6. The abandonment rate in the $GI/GI/N+GI$ queue with two classes, each having Poisson arrivals with rate 60, $N = 100$ servers, and $LN(1,v)$ patience distribution. The predicted class 1 abandonment rate is 0 and that for class 2 is $(\lambda_2 - b_2\mu_2) \times N = 20$.

The static priority scheduling policy π_{sp} does not depend on the patience distributions and only depends on the service time distributions through their means. In contrast, the objective function in (1) depends on the full details of the service and patience distributions. This raises the question as to how well the fluid invariant states approximate the steady-state behavior of the $GI/GI/N+GI$ queue operating under π_{sp}. Figure 6[5] shows numerically that neither the patience distribution nor the service time distribution have much impact on the rate at which customers abandon. In Figure 6, the two service distributions considered

[5] Each simulation shown in Figure 6, and later in Tables 1 and 2, is run until 5 million customers arrive, with the time before the arrival of the 1000th customer considered a "warm-up" period and discarded. The numbers graphed in Figure 6 are the average over 10 runs. The difference between the largest and smallest of the 10 run values for each class 2 value graphed is less than 1%. (For class 1, the absolute difference is very small, but the percentage can be larger when the variance is small, due to the values being very close to 0.)

are Lognormal(1,4) and Gamma(2)[6] defined at the end of Section 4, which both have mean 1, but have variance 4 and 0.5 respectively, and the patience distributions are those considered in Figure 5.

The class 2 mean steady-state queue-length does depend on the patience distribution under the static priority scheduling policy that assigns class 1 higher priority than class 2, which is in contrast to the abandonment percentage. (The class 1 queue-length is close to zero, due to its priority, and so does not depend on any of the primitive distributions.) The dependence on the queue-length is not surprising given Theorem 1. Table 1 shows that the function q_j defined in Theorem 1 for $j \in \mathbb{J}$ provides a good approximation of the class 2 queue-lengths. Specifically, for the system in Table 1, if we recall Remark 8 and write all parameters in terms of $N = 100$, then we can see that the system has class j arrival rate $0.6N$, so that the fluid arrival rate is $\lambda_j = 0.6$ for $j \in \{1,2\}$ and the class j workload contribution is $\rho_j = 0.6$ ($\mu_j = 1$ for $j \in \{1,2\}$ because the mean service time for both classes is 1). The proportion of server effort devoted to class 1 is exactly enough to handle all the arrivals, meaning $b_1 = \lambda_1 = 0.6$. The remainder is devoted to class 2, meaning $b_2 = 1 - 0.6 = 0.4$. Both classes have patience distributions with mean 1, which implies $\theta_1 = \theta_2 = 1$. Then, if Q_1^N and Q_2^N are random variables that represent the steady-state queue-lengths for each class, we have the approximations

$$\mathbb{E}\left[Q_1^N\right] \approx N q_1(0.6) = 0 \text{ and } \mathbb{E}\left[Q_2^N\right] \approx N q_2(0.4) = 0.6 G_{e,j}^r \left(\left(G_j^r\right)^{-1} \left(1 - \frac{0.4}{0.6}\right)\right).$$

	Predicted		Approximated	
v	$\mathbb{E}[Q_1^N]$	$\mathbb{E}[Q_2^N]$	$\mathbb{E}[Q_1^N]$	$\mathbb{E}[Q_2^N]$
0.2	0	42.1120	1.5215 +/- 0.0043	39.9417 +/- 0.0637
0.5	0	33.3982	1.5130 +/- 0.0038	31.9239 +/- 0.0665
1	0	25.9555	1.4904 +/- 0.0048	25.2024 +/- 0.0789
2	0	18.8790	1.4252 +/- 0.0048	18.6306 +/- 0.0757
5	0	11.4800	1.2600 +/- 0.0054	11.5150 +/- 0.0609

Table 1. A comparison of the queue-length approximations given in Theorem 4.2 with their simulated steady-state mean (95% confidence intervals shown) in a $M/LN(1,4)/100 + LN(1,v)$ queue with two classes, each having Poisson arrivals with rate 60.

The reader can observe in Table 1 that the queue-lengths decrease as the variability of the patience distribution increases, which is consistent with the theoretic prediction in Figure 5. This effect becomes even more pronounced when the patience distribution is Gamma(p), as can be seen in Table 2. One contributing factor may be that the Gamma(p) distribution has an associated hazard rate function that is strictly increasing when $p > 1$ and strictly decreasing when $p < 1$, whereas the hazard rate function associated with a lognormal distribution is not monotonic. Table 2 also shows that the prediction error is larger when the hazard rate function is strictly decreasing rather than when it is strictly increasing. Further simulations (not shown) confirm the prediction error does decrease as the system size becomes larger, consistent with the underlying theory in [59]. However, we do not know the connection between the rate at which the prediction error becomes small and the patience distribution characteristics except in a single class $M/GI/N + GI$ setting; see [11, Theorem 5].

[6] The service time distributions we assume in Figure 6 are the ones used in Table 1 in [64] in the single class setting, to illustrate the lack of effect of the service distribution on steady-state system performance measures for overloaded systems.

	Predicted		Approximated	
p	$\mathbb{E}[Q_1]$	$\mathbb{E}[Q_2]$	$\mathbb{E}[Q_1]$	$\mathbb{E}[Q_2]$
5	0	41.5688	1.5188 +/- 0.0043	39.2277 +/- 0.0491
2	0	30.8627	1.4863 +/- 0.0053	29.4272 +/- 0.1026
1	0	20.0000	1.3280 +/- 0.0032	18.8971 +/- 0.0718
0.5	0	8.6276	0.9225 +/- 0.0020	7.4433 +/- 0.0283
0.2	0	0.5831	0.3352 +/- 0.0012	1.0432 +/- 0.0073

Table 2. A comparison of the fluid queue approximations given in Theorem 4.2 with their simulated steady-state mean (95% confidence intervals shown) in a $M/LN(1,4)/100+\text{Gamma}(p)$ queue with two classes, each having Poisson arrivals with rate 60.

Remark 9. (Consistency Check) We can double-check that the simulation results shown in Figure 1 are consistent with theory. As in the examples earlier in this section, the class j arrival rate is $0.6N$, so that the fluid arrival rate is $\lambda_j = 0.6$ for $j \in \{1,2\}$, the class j workload contribution is $\rho_j = 0.6$, the proportion of server effort devoted to class 1 is $b_1 = \lambda_1 = 0.6$, and the proportion of server effort devoted to class 2 is $b_2 = 1 - b_1 = 0.4$. Substituting into the formulae (25) and (26), and noting that $j^* = 2$ in this example, show that

$$q_2(b_2) = \frac{\lambda_2 - b_2\mu_2}{\theta_2} = 0.6 - 0.4 \times 1 = 0.2,$$

when the class 2 patience distribution is exponential with mean 1, and

$$q_2(b_2) = \frac{\lambda_2}{\theta_2}\left(1 - \left(\frac{b_2}{\rho_2}\right)^2\right) = 0.6 \times \left(1 - \left(\frac{0.4}{0.6}\right)^2\right) = 1/3,$$

when the class 2 patience distribution is uniform(0,2). Multiplying both of the above numbers by $N = 100$ approximately matches the queue-lengths shown in Figure 1 under the static priority policy that assigns priority to class 1.

7. Open Problems

Holding Costs

The scheduling problem (1) does not use holding costs to penalize congestion. However, the scheduling policy known as the $c\mu$ rule was first introduced in [62] in a setting with holding costs and no customer abandonment (the c in $c\mu$ refers to the holding cost vector), and many works use holding costs to penalize congestion. Holding costs are a natural way to capture the disutility associated with waiting, and, in our setting, do this in a less extreme manner than abandonment costs. This motivates us to suppose there is an additional class-dependent holding cost $c_j, j \in \mathbb{J}$, incurred per customer per unit time. Then, if $Q_j^N(t,\pi^N), j \in \mathbb{J}$, represents the number of class j customers waiting in queue at time $t \in \mathbb{R}_+$ under scheduling policy $\pi \in \Pi$, the modified scheduling problem objective function is

$$\mathcal{C}^N(\pi^N) := \limsup_{T\to\infty} \frac{1}{T} I\!E\left[\sum_{j=1}^{J} a_j R_j^N(T,\pi^N) + \int_0^T c_j Q_j^N(t,\pi^N)dt\right]. \tag{27}$$

The modified scheduling problem is easily translated to a modified fluid control problem. Specifically, from Theorem 1, when the server effort allocation vector is $b \in \mathbb{B}$, then $q_j(b)$ gives the associated class $j \in \mathbb{B}$ fluid queue. This leads to the modified fluid control problem

$$\inf_{b\in\mathbb{B}} \sum_{j=1}^{J} a_j(\lambda_j - b_j\mu_j) + c_j q_j(b_j). \tag{28}$$

Suppose the patience distribution is exponential. Then, similar to (25), $q_j(b) = (\lambda_j - b\mu_j)/\theta_j$ for $j \in \mathbb{J}$ and $b \in [0, \rho_j]$, and so (28) becomes

$$\inf_{b\in\mathbb{B}} \sum_{j=1}^{J} a_j\left(\lambda_j - b_j\mu_j\right) + c_j \frac{\lambda_j - b_j\mu_j}{\theta_j}$$

Define the modified cost that incorporates both holding and abandonment

$$\tilde{c}_j = a_j + \frac{c_j}{\theta_j}, \text{ for all } j \in \mathbb{J}$$

and observe that (28) becomes

$$\inf_{b\in\mathbb{B}} \sum_{j=1}^{J} \tilde{c}_j\left(\lambda_j - b_j\mu_j\right). \tag{29}$$

In particular, the solution to the LP (28) is exactly as in Section 6, except with $\tilde{c}_j$ replacing c_j for $j \in \mathbb{J}$, as observed in [6]. This suggests that the static priority $\tilde{c}\mu$ scheduling policy that ranks classes in the order of their $\tilde{c}_j\mu_j$ values should approximately solve (27).

The issue is that the solution to (28) is in general more complicated. However, a strong simplification occurs when the patience distributions all have strictly increasing hazard rate functions. For example, the Gamma distributions plotted in Figure 5(a) have strictly increasing hazard rate functions when $p > 1$.

Lemma 1. *Assume g_j^r is positive and continuous on $(0, H_j^r)$ for all $j \in \mathbb{J}$. If h_j^r is (strictly) increasing on $(0, H_j^r)$, then q_j is (strictly) concave on $(0, \rho_j)$.*

Lemma 1, proved in Section 9, implies that the objective function in (28) is concave. Then, the solution occurs at a feasible vertex, meaning that the solution b^* has $b_j^* \in \{0, \rho_j\}$ for $J-1$ of the classes and b^* determined so that $\sum_{j=1}^{J} b_j^* = 1$, under the condition (24). This would be the solution in Section 6, *except* that how to order the classes is not clear, except in special cases (for example, the calculation in (26) can be extended to $J > 2$).

Remark 10. (Non-Optimality of Static Priority) We do not wish to leave the impression that the modified fluid control problem (28) always has solutions that occur at a feasible vertex. The reader that wishes to do additional work will find that Lemma 1 can be modified to show that when h_j^r is (strictly) decreasing, then q_j is (strictly) convex on $(0, \rho_j)$. The implication is that the solution may occur at an interior point, which motivates the need to study scheduling policies that have full flexibility to partially serve classes (i.e., the need for scheduling policies more general than static priority).

Asymptotic Optimality

We would like to rigorously establish that the (modified) fluid control problem (28) arises as the limit of the scheduling problem (27) when the number of servers becomes large and the arrival process satisfies a functional strong law like (4). More specifically, if $\mathcal{C}^N(\pi)$ is the long run average cost when the number of servers is N, we would like to show that the class of scheduling policies Π is such that

$$\lim_{N\to\infty} \frac{C^N(\pi)}{N} \geq m^*, \text{ for all } \pi \in \Pi, \tag{30}$$

recalling that m^* is the minimum objective function value for the fluid control problem in (23) (or in (28) when holding costs are positive). For $GI/GI/N+M$ systems (that is, when the patience distributions are all exponential), (30) has been shown to hold under mild asymptotic conditions; see [7, Proposition 2.1 and A.1]. This leaves open the question of showing (30) when the patience distribution is not assumed to be exponential, which is work in progress of the authors of this tutorial paper.

We would like to further define a scheduling policy $\pi^* \in \Pi$ such that

$$\lim_{N\to\infty} \frac{C^N(\pi^*)}{N} = m^*. \tag{31}$$

In the case that holding costs are 0, $\pi^* = \pi_{sp}$ is the static priority scheduling policy defined in Section 6. For $GI/GI/N+M$ systems, [9, Theorem 5.1] and its proof provide conditions under which (31) holds for the static priority policy determined from the solution to (29), as explained towards the end of Section 6. These conditions include natural asymptotic assumptions, and require that the system steady-state, under fluid scaling, converge to the unique invariant state[7]. Work along these lines for more general patience distributions is also work in progress of the authors of this tutorial paper.

Joint Staffing and Scheduling

The fluid control problems (23) and (28) that approximate the scheduling problems (1) and (27) are non-trivial only when the system is overloaded in the sense that

$$\sum_{j=1}^{J} \frac{\lambda_j^N}{\mu_j} > N, \tag{32}$$

recalling λ_j^N is the class j arrival rate to the multiclass $G/GI/N+GI$ queue given in Section 2 when the arrival rate does not vary with time. The inequality (32) is exactly the condition (24) used when solving the fluid control problem (23) if, for example, the arrival processes $E_j^N, j \in \mathbb{J}$, are all renewal processes with rate $\lambda_j^N = \lambda_j N$ for a specified $\lambda_j \in \mathbb{R}_+$.

The issue is that the fluid control problem *assumes* the decision has been made to not staff enough to serve all the customers. This can be cost effective. For example, for a system with one customer class ($J=1$) a large arrival rate, and linear staffing, holding, and abandonment costs, a sufficient condition to ensure that a minimum cost staffing decision results in an overloaded system (i.e., (24) holds) is that the customer patience distribution has decreasing hazard rate; see [11, Proposition 4]. The decreasing hazard rate implies that customers become less and less likely to abandon as they wait, modeling a situation in which the time invested in waiting increases the customer commitment to receive the service. In other words, taking advantage of customer-willingness-to-wait is advantageous.

We would like to understand the solution to the joint staffing and scheduling problem; that is, the solutions to (1) and (27), when the number of servers is also a decision. A related problem for a multiclass system with no abandonment and constraints on customer waiting is solved in [30]. Another related joint staffing and routing problem in a system with exponentially distributed customer patience times, one customer class, and multiple server pools is solved in [2]. However, no joint staffing and scheduling problem has been solved when the customer patience distributions are not exponential. In many practical settings, customer patience times are not exponentially distributed (as was verified in the call center setting in [15]). Therefore, one key question of interest is to determine how sensitive the solution structure is to the exponential assumption.

[7] For this, the authors cite [9, Theorem 4.4], the proof of which is incomplete due to its reliance on the proof techniques of [43, Theorem 3.3], and an erratum is currently in progress.

Time-Varying Arrival Rates

The fluid model in Section 3 allows for time-varying input. Moreover, the fluid model can be modified to many-server systems in which customers do not abandon, as detailed in Remark 5 (and when $J=1$ the relevant fluid model and supporting convergence result can be found in [44]). This implies that the fluid model in Section 3 is potentially relevant for a wide variety of application environments, whenever many-server models are relevant. For example, many-server models have been used to model data centers [47, 25], call centers [1, 26], and hospital operations [16, 37, 24, 17], all of which experience time-varying demand.

The static priority $a\mu$-rule (more commonly known as the $c\mu$-rule), that is shown to solve the fluid control problem in Section 6, is very appealing because it is simple, easy to implement, and has been shown to be optimal or asymptotically optimal in a wide variety of settings, beginning with the early work of [19, 62], continuing with the later work of [56] (which provides an excellent literature review), [53], and this paper. However, we are not aware of any optimality results for the $a\mu$-rule when customer arrivals are time-varying (although [36] discusses how to make control decisions to stabilize performance measures for different classes). We are hopeful that the fluid model presented in Section 3 can be used to formulate a control problem when customer arrivals are time-varying that has the static priority $a\mu$-rule as its solution.

Even better would be to be able to use the knowledge that a static priority scheduling rule performs well to jointly determine scheduling and staffing in a time-varying setting. There is work on determining staffing levels when there is time-varying demand in the $J=1$ setting, as in [28, 34, 48, 50], and related papers discuss how to calculate fluid performance measures [49], and how to develop wait-time predictors [40]. However, we are not aware of any work that jointly considers staffing and scheduling.

Connection to Diffusion Approximations

We focus on the prelimit system described in Section 2, and drop the superscript N in this subsection to emphasize that. Then, consistent with (32), the fluid control problems (23) and (28) are non-trivial only when the system is overloaded in the sense that

$$\sum_{j=1}^{J}\frac{\lambda_j}{\mu_j} > N.$$

Assuming N, λ_j, and μ_j for all $j\in\mathbb{J}$ are known parameters, that were, for example estimated from data, one way to quantify the amount of system overload is to solve for β in

$$N=\sum_{j=1}^{J}\frac{\lambda_j}{\mu_j}+\beta\sqrt{\sum_{j=1}^{J}\frac{\lambda_j}{\mu_j}},\ \text{ for } \beta\in\mathbb{R}_+.$$

The size of β can help us determine whether the fluid control problems in (23) and (28) are most relevant, or whether a different approximation to the scheduling control problems in (1) and (27) will be more helpful. For example, a small value of β suggests that we may want to study the approximating control problem that arises in the regime known as the quality-and-efficiency (QED) driven regime, first introduced in [31] for a single class ($J=1$) many-server system without abandonment and later expanded to include abandonment when patience distributions are exponential in [27], and to non-exponential distributions in [20, 23, 52, 60, 66]; for a comprehensive review of the QED regime see the tutorial papers [21, 46], for very recent work in the QED regime see the large deviations analysis in [57], and for an overview of the different regimes see the survey papers [63, 22]. In multiclass systems ($J>1$), the papers [33, 45] study scheduling control problems in the QED regime. Similarly, we may also want to study the approximating control problem that arises in the regime known as the non-degenerate slowdown (NDS) regime; see [4] for the development of this regime, and [5, 8, 10, 29] for some example control problems in that regime. In contrast to the QED regime, the NDS regime so far has only been developed for systems without abandonment, and so a first step would be to generalize that regime to include customer impatience. Next, we may want to investigate the ED+QED regime, developed in [54] as a refinement to the efficiency-driven (ED) regime in which fluid models are relevant (because the system is overloaded), and also studied in [23, 39]. We end by observing that neither the QED regime, nor the NDS regime, nor the QED+ED regime appears able to handle fully general systems (that is, systems with non-exponential interarrival, service, and abandonment distributions), as we do here.

The recent work [14] builds on ideas in [3, 38] to provide a step in the direction of unifying the approximating control problems that arise in the different regimes when all primitive input distributions are exponential. The question of how to quantify the trade-offs of using the different regimes to develop an approximating scheduling control problem for one given system is challenging. The issue is that the different approximating control problems can potentially motivate different scheduling policies.

8. Proof of Theorem 1

Let

$$f_j(x) := \begin{cases} 1, & \text{if } x = 1, \\ G^r_{e,j}\left(\left(G^r_j\right)^{-1}(x)\right), & \text{if } x \in [0,1), \end{cases} \quad \text{for } j \in \mathbb{J}.$$

Then, by (17),

$$q_j(b_j) = \begin{cases} \frac{\lambda_j}{\theta_j}, & \text{if } b_j = 0, \\ \frac{\lambda_j}{\theta_j} f_j\left(1 - \frac{b_j}{\rho_j}\right), & \text{if } b_j \in (0, \rho_j], \end{cases} \quad \text{for } j \in \mathbb{J}.$$

We first show that a tuple satisfying conditions (i), (ii), and (iii) in Theorem 1 is an invariant state, as given in Definition 1. We second show the converse.

Proof of forward direction. Fix $b \in \mathbb{B}$ and let $(X^*, \nu^*, \eta^*) \in \mathbb{X}$ be defined by (i)-(iii) in Theorem 1. Let (X, ν, η) be the constant function such that $(X(t), \nu(t), \eta(t)) = (X^*, \nu^*, \eta^*)$ for all $t \geq 0$. We must show (X, ν, η) is a fluid model solution; i.e., that $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$, that $B, Q, R, D, K \in \mathbf{C}(\mathbb{R}^J_+)$, and that (5), (6), and (a)-(d) in Definition 1 hold. First observe that (i) and (ii) imply that

$$\left\langle h^r_j, \eta^*_j \right\rangle = \int_0^{H^r_j} h^r_j(x) \lambda_j \left(1 - G^r_j(x)\right) dx = \lambda_j < \infty, \ \text{for } j \in \mathbb{J},$$

and

$$\left\langle h^x_j, \nu^*_j \right\rangle = \int_0^{H^s_j} h^s_j(x) b_j \mu_j \left(1 - G^s_j(x)\right) dx = b_j \mu_j < \infty, \ \text{for } j \in \mathbb{J},$$

and so (5) holds. Next, to see the desired continuity and that (a)-(d) in Definition 1 hold, we begin by observing that, for each $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$,

$$B_j(t) = B^*_j = b_j, \ \text{by (ii) and (7)}, \tag{33}$$

$$Q_j(t) = Q^*_j = \frac{\lambda_j}{\theta_j} f_j(1 - b_j/\rho_j), \ \text{by (iii) and (8)}, \tag{34}$$

$$D_j(t) = b_j \mu_j t, \ \text{by (ii) and (11)}, \tag{35}$$

$$K_j(t) = b_j \mu_j t, \ \text{by (12)}. \tag{36}$$

From (33)-(36), $B, Q, D, K \in \mathbf{C}(\mathbb{R}^J_+)$, and from (i)-(iii) $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$. The fact that $R \in \mathbf{C}(\mathbb{R}^J_+)$ follows from our argument below that (b) holds, which requires showing (13) holds.

Equation (33) implies (a), because $b \in \mathbb{B}$. Equation (36) shows K_j is non-decreasing for each $j \in \mathbb{J}$, meaning (c) holds.

In this paragraph, we argue that (b) holds. The inequality $0 \le Q_j(t) \le \langle 1, \eta_j \rangle$ for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$ follows from (34), because f_j has range $[0,1]$ and $\langle 1, \eta_j \rangle = \lambda_j/\theta_j$ for all $j \in \mathbb{J}$ from (i). Next, for each $j \in \mathbb{J}$, let χ_j be the unique solution of

$$G^r_{e,j}(\chi_j) = f_j(1 - b_j/\rho_j) \text{ if } b_j > 0 \text{ and } H^r_j \text{ otherwise.}$$

Then, letting $\left(G^r_{e,j}\right)^{-1}$ be the inverse function for $G^r_{e,j}$,

$$\chi_j = \left(G^r_{e,j}\right)^{-1}\left(f_j(1 - b_j/\rho_j)\right) = \left(G^r_j\right)^{-1}\left(1 - b_j/\rho_j\right), \text{ for each } j \in \mathbb{J} \text{ such that } b_j > 0,$$

and so by (i), (9), and (10),

$$\begin{aligned} R_j(t) &= \lambda_j \int_0^t \int_0^{\chi_j} g^r_j(x)dx \\ &= \begin{cases} \lambda_j t, & \text{if } b_j = 0, \\ \lambda_j G^r_j(\chi_j)t, & \text{if } b_j \in (0, \rho_j], \end{cases} \\ &= \begin{cases} \lambda_j t, & \text{if } b_j = 0, \\ (\lambda_j - b_j\mu_j)\,t, & \text{if } b_j \in (0, \rho_j], \end{cases} \end{aligned} \tag{37}$$

for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$. By (37) and the expression for D_j in (35), $D_j(t) + R_j(t) = \lambda_j t = E_j(t)$ for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$, and so (X, ν, η) satisfies (13). Thus, (b) in Definition 1 holds.

In this paragraph, we argue that (d) in Definition 1 holds; i.e., that (14) and (15) are satisfied. By (ii) and (36), for any continuous and bounded function f having domain $\mathbb{R}_+$, for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$,

$$\begin{aligned} &\int_0^\infty \frac{1 - G^s_j(x+t)}{1 - G^s_j(x)} f(x+t)\nu^*_j(dx) + \int_0^t (1 - G^s_j(t-u))f(t-u)dK_j(u) \\ &= b_j\mu_j \int_0^\infty \left(1 - G^s_j(x+t)\right) f(x+t)dx + b_j\mu_j \int_0^t (1 - G^s_j(u))f(u)du \\ &= b_j\mu_j \int_0^\infty \left(1 - G^s_j\right)(u)du \\ &= \left\langle f, \nu^*_j \right\rangle. \end{aligned}$$

Since $\nu_j(t) = \nu^*_j$ for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$, (14) holds. Finally, by (i), for any continuous and bounded function f having domain $\mathbb{R}_+$, for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$,

$$\int_0^\infty \frac{1 - G^r_j(x+t)}{1 - G^r_j(x)} f(x+t)\eta^*_j(dx) + \lambda_j \int_0^t (1 - G^r_j(t-s))f(t-s)ds = \left\langle f, \eta^*_j \right\rangle.$$

Since $\eta_j(t) = \eta^*_j$ for all $j \in \mathbb{J}$ and $t \in \mathbb{R}_+$, (15) holds.

In summary, we have established $(X, \nu, \eta) \in \mathbf{C}(\mathbb{X})$ satisfies (5) and (6), and that $B, Q, R, D, K \in \mathbf{C}(\mathbb{R}^J_+)$ satisfy conditions (a)-(d). We conclude $(X^*, \nu^*, \eta^*) \in \mathcal{I}_\lambda$.

Proof of converse direction. Suppose that $(X^*, \nu^*, \eta^*) \in \mathcal{I}_\lambda$. Let (X, ν, η) be the constant function such that $(X(t), \nu(t), \eta(t)) = (X^*, \nu^*, \eta^*)$ for all $t \in \mathbb{R}_+$. By assumption (X, ν, η) is a fluid model solution. Therefore, from Definition 1(a), $0 \le B^*_j \le 1$. In order to show that $B^* \in \mathbb{B}$, we must also show that $B^*_j \le \rho_j$ for each $j \in \mathbb{J}$. We will verify this in the process of verifying that (X^*, ν^*, η^*) satisfies (i)-(iii) in Theorem 1 for $b = B^*$.

We first show (i) of Theorem 1 holds. By (15), for f that is continuous and bounded and has domain $\mathbb{R}_+$, $j \in \mathbb{J}$, and $t \in \mathbb{R}_+$,

$$\langle f, \eta^*_j \rangle = \int_0^\infty \frac{1 - G^r_j(x+t)}{1 - G^r_j(x)} f(x+t)\eta^*_j(dx) + \lambda_j \int_0^t (1 - G^r_j(t-u))f(t-u)du.$$

For each $j \in \mathbb{J}$, as $t \to \infty$, the first integral converges to zero by dominated convergence, and the second to $\lambda_j \int_0^\infty (1 - G_j^r(u)) f(u) du$. Since f was an arbitrary continuous and bounded function on $\mathbb{R}_+$, (i) of Theorem 1 holds.

We next show (ii) of Theorem 1 holds. From Definition 1(b) and the fact that we have already established (i) of Theorem 1, $0 \le Q_j^* \le \langle 1, \eta_j^* \rangle = \lambda_j / \theta_j$ for each $j \in \mathbb{J}$. Define $\chi_j \in [0, H_j^r]$ to be the unique solution to

$$Q_j^* = \int_0^{\chi_j} \eta_j^*(u) du, \text{ for each } j \in \mathbb{J}.$$

so that χ_j satisfies (9). We fix $j \in \mathbb{J}$ and separate the cases that $\chi_j = H_j^r$ and $\chi_j < H_j^r$.

Case $\chi_j = H_j^r$. From (i) and (10),

$$R_j(t) = \int_0^t \int_0^{H_j^r} h_j^r(w) \lambda_j (1 - G_j^r(w)) dw = \lambda_j t, \text{ and all } t \in \mathbb{R}_+.$$

From (13),

$$D_j(t) = E_j(t) - R_j(t) = 0, \text{ and all } t \in \mathbb{R}_+,$$

and from (12),

$$K_j(t) = 0, \text{ and all } t \in \mathbb{R}_+.$$

From (14), for all continuous and bounded f with domain $\mathbb{R}_+$,

$$\langle f, \nu_j^* \rangle = \int_0^{H_j^s} \frac{1 - G_j^s(x+t)}{1 - G_j^s(x)} f(x+t) \nu_j^*(dx) \to 0 \text{ as } t \to \infty, \tag{38}$$

noting that the dominated convergence theorem validates the limit in (38). Since (38) implies $\langle f, \nu_j^* \rangle = 0$ for all continuous and bounded f with domain $\mathbb{R}_+$, from (7), $B_j^* = 0$, and so (ii) holds trivially.

Case $\chi_j < H_j^r$. From (i) and (9), $Q_j^* = \lambda_j G_{e,j}^r(\chi_j)/\theta_j$, or, equivalently, $\chi_j = \left(G_{e,j}^r\right)^{-1}(\theta_j Q_j^* / \lambda_j)$. From (i) and (10),

$$R_j(t) = \int_0^t \left(\int_0^{\chi_j} h_j^r(w) \lambda_j (1 - G_j^r(w)) dw \right) du = \lambda_j G_j^r \left(\left(G_{e,j}^r\right)^{-1} (\theta_j Q_j^* / \lambda_j) \right) t,$$

for all $t \in \mathbb{R}_+$. Set

$$k_j^* = \lambda_j \left(1 - G_j^r \left(\left(G_{e,j}^r\right)^{-1} (\theta_j Q_j^* / \lambda_j) \right) \right).$$

From (13),

$$D_j(t) = k_j^* t, \text{ for all } t \in \mathbb{R}_+,$$

and from (12),

$$K_j(t) = k_j^* t, \text{ for all } t \in \mathbb{R}_+.$$

From (14), for all continuous and bounded f with domain $\mathbb{R}_+$,

$$\langle f, \nu_j^* \rangle = \int_0^\infty \frac{1 - G_j^s(x+t)}{1 - G_j^s(x)} f(x+t) \nu_j^*(dx) + k_j^* \int_0^t (1 - G_j^s(t-u)) f(t-u) du, \tag{39}$$

for all $t \in \mathbb{R}_+$. As $t \to \infty$, the first integral in (39) converges to 0, by dominated convergence, as in (38), and the second to $k_j^* \int_0^\infty (1 - G_j^s(u)) f(u) du$. Hence

$$\nu_j^*(du) = k_j^* (1 - G_j^s(u)) du \text{ for all } u \in \mathbb{R}_+.$$

From (7), $B_j^* = k_j^* / \mu_j$, and so (ii) of Theorem 1 holds for $b_j = B_j^*$.

The fact that $B^* \in \mathbb{B}$ follows because for any $j \in \mathbb{J}$ such that case (i) holds, $B_j^* = 0$, and for any $j \in \mathbb{J}$ such that case (ii) holds, $k_j^* \le \lambda_j$ implies $B_j^* \le \rho_j$.

Finally, to see (iii) of Theorem 1 holds, observe that from the argument used to establish (ii) above, for any $j \in \mathbb{J}$, if $\chi_j = H_j^r$, then $B_j^* = 0$ and

$$Q_j^* = \lambda_j/\theta_j.$$

Otherwise, for $j \in \mathbb{J}$ such that $\chi_j < H_j^r$, the definition of k_j^* and the fact that $B_j^* = k_j^*/\mu_j$ implies

$$B_j^* = \frac{\lambda_j}{\mu_j}\left(1 - G_j^r\left(\left(G_{e,j}^r\right)^{-1}\left(\frac{\theta_j Q_j^*}{\lambda_j}\right)\right)\right).$$

Re-arranging terms in the above equality shows that for such a $j \in \mathbb{J}$,

$$Q_j^* = \frac{\lambda_j}{\theta_j} G_{e,j}^r\left(\left(G_j^r\right)^{-1}\left(1 - \frac{B_j^*}{\rho_j}\right)\right).$$

We conclude Q_j^* and B_j^* satisfy (17) for all $j \in \mathbb{J}$, which implies (iii) by (8).

9. Proof of Lemma 1

Fix $j \in \mathbb{J}$. By assumption, g_j^r is positive and continuous on $(0, H_j^r)$. In particular, by the fundamental theorem of calculus, G_j^r is continuously differentiable on $(0, H_j^r)$ with positive derivative g_j^r. Then, by the inverse function theorem, $(G_j^r)^{-1}$ is differentiable at each $x \in (0,1)$, and

$$\frac{d}{dx}(G_j^r)^{-1}(x) = \frac{1}{g_j^r\left((G_j^r)^{-1}(x)\right)}.$$

Also, by the fundamental theorem of calculus, $G_{e,j}^r$ is continuously differentiable on $(0, H_j^r)$ with derivative

$$\frac{d}{dx} G_{e,j}^r(x) = \theta_j(1 - G_j^r(x)), \text{ for } x \in (0, H_j^r).$$

We use the chain rule in (17) to find that for $b \in (0, \rho_j)$

$$q_j'(b) = -\mu_j \frac{\left(1 - G_j^r\left(\left(G_j^r\right)^{-1}(1 - b/\rho_j)\right)\right)}{g_j^r\left(\left(G_j^r\right)^{-1}(1 - b/\rho_j)\right)} = \frac{-\mu_j}{h_j^r\left(\left(G_j^r\right)^{-1}(1 - b/\rho_j)\right)}.$$

As $b \in (0, \rho_j)$ increases, $1 - b/\rho_j$ decreases, causing $\left(G_j^r\right)^{-1}(1 - b/\rho_j)$ to decrease and in turn $h_j^r\left(\left(G_j^r\right)^{-1}(1 - b/\rho_j)\right)$ to decrease. Hence, $q_j'(b)$ is decreasing in $b \in (0, \rho_j)$, and so q_j is concave on $(0, \rho_j)$. When h_j^r is strictly increasing, then q_j is strictly concave on $(0, \rho_j)$.

Acknowledgememnts

We thank Ohad Perry (Northwestern University, Faculty) for reading a preliminary draft and offering valuable comments. We thank Marvin Pena (California State University San Marcos, Undergraduate) for deriving the fluid queue invariant state in (26). We thank Huiyu Wu (The University of Chicago, Masters) for programming the code required to produce Figures 1, 5, and 6, and Tables 1 and 2. We thank James Kiselik (The University of Chicago, Booth Faculty Services) for a thorough reading of the paper and for many editorial improvements.

References

[1] Z. Aksin, M. Armony, and V. Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management*, 16(6):665–688, 2007.

[2] M. Armony and A. Mandelbaum. Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Operations Research*, 59(1):50–65, 2011.

[3] B. Ata and I. Gurvich. On optimality gaps in the Halfin-Whitt regime. *Annals of Applied Probability*, 22(1):407–455, 2012.

[4] R. Atar. A diffusion regime with nondegenerate slowdown. *Operations Research*, 60(2):490–500, 2012.

[5] R. Atar and A. Cohen. Asymptotically optimal control for a multiclass queueing model in the moderate deviation heavy traffic regime. *Annals of Applied Probability*, 27(5):2862–2906, 2017.

[6] R. Atar, C. Giat, and N. Shimkin. The $\frac{c\mu}{\theta}$ rule for many-server queues with abandonment. *Operations Research*, 58(5):1427–1439, 2010.

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

[8] R. Atar and I. Gurvich. Scheduling parallel servers in the non-degenerate slowdown diffusion regime: Asymptotic optimality results. *Annals of Applied Probability*, 24(2):760–810, 2014.

[9] R. Atar, H. Kaspi, and N. Shimkin. Fluid limits for many-server systems with reneging under a priority policy. *Mathematics of Operations Research*, 39(3):672–696, 2014.

[10] R. Atar and S. Saha. Optimality of the generalized $c\mu$ rule in the moderate deviation regime. *Queueing Systems*, 87(1-2):113–130, 2017.

[11] A. Bassamboo and R. S. Randhawa. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research*, 58(5):1398–1413, 2010.

[12] A. Bassamboo and R. S. Randhawa. Scheduling homogeneous impatient customers. *Management Science*, 62(7):2129–2147, 2016.

[13] P. Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, Inc., 1999. Second Edition.

[14] A. Braverman, I. Gurvich, and J. Huang. On the Taylor expansion of value functions, 2018. arXiv:1804.05011v1.

[15] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the Americal Statistical Association*, 100(469):36–50, 2005.

[16] C. W. Chan, J. Dong, and L. V. Green. Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research*, 65(2):469–495, 2017.

[17] C. W. Chan and V. Sarhangian. Dynamic server assignment in multiclass queues with shifts, with applications to nurse staffing in emergency departments, 2019. Working Paper.

[18] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks*. Springer-Verlag, 2001.

[19] D. Cox and W. Smith. *Queues*. Methuen, London, 1961.

[20] J. G. Dai and S. He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, 2010.

[21] J. G. Dai and S. He. Queues in service systems: Customer abandonment and diffusion approximations. In J. Geunes, editor, *Tutorials in Operations Research*, pages 36–59. INFORMS, 2011.

[22] J. G. Dai and S. He. Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering*, 21:1–36, 2012.

[23] J. G. Dai, S. He, and T. Tezcan. Many-server diffusion limits for $\mathrm{G}/\mathrm{P}h/n+\mathrm{GI}$ queues. *Annals of Applied Probability*, 20(5):1854–1890, 2010.

[24] J. Dong and O. Perry. Queueing models for patient-flow dynamics in inpatient wards, 2019. Forthcoming in *Operations Research.*

[25] A. Gandhi, M. Harchol-Balter, R. Das, and C. Lefurgy. Optimal power allocation in server farms. *Performance Evaluation Review*, 37(1):157–169, 2009. Special Issue: Proceedings of the ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems, June 15-19.

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

[27] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manfacturing & Service Operations Management*, 4:208–227, 2002.

[28] L. Green, P. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

[29] V. Gupta and N. Walton. Load balancing in the non-degenerate slowdown regime, 2018. arXiv:1707.01969v2.

[30] I. Gurvich, M. Armony, and A. Mandelbaum. Service-level differentiation in call centers with fully flexible servers. *Management Science*, 54(2):279–294, 2008.

[31] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, 1981.

[32] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action.* Cambridge University Press, 2013.

[33] J. M. Harrison and A. Zeevi. Dynamic scheduling of a multiclass queue in the Halfin-Whitt heavy traffic regime. *Operations Research*, 51:243–257, 2004.

[34] B. He, Y. Liu, and W. Whitt. Staffing a service system with non-Poisson nonstationary arrivals. *Probability in the Engineering and Informational Sciences*, 30(4):593–621, 2016.

[35] W. Hopp and W. Lovejoy. *Hospital Operations: Principles of High Efficiency Health Care.* Financial Times Press, 2013.

[36] K. Hovey, Y. Liu, and X. Sun. Staffing and scheduling in time-varying multiclass service systems to achieve service level differentiation, 2018. Working Paper.

[37] J. Huang, B. Carmeli, and A. Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):751–978, 2015.

[38] J. Huang and I. Gurvich. Beyond heavy-traffic regimes: Universal bounds and controls for the single-server queue. *Operations Research*, 66(4):1168–1188, 2018.

[39] J. Huang, A. Mandelbaum, H. Zhang, and J. Zhang. Refined models for efficiency-driven queues with applications to delay announcements and staffing. *Operations Research*, 65(5):1380–1397, 2017.

[40] R. Ibrahim and W. Whitt. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, 59(5):1106–1118, 2011.

[41] W. Kang and G. Pang. Equivalence of fluid models for $\mathrm{G}_t/\mathrm{GI}/\mathrm{N}+\mathrm{GI}$ queues, 2019. Working Paper.

[42] W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Annals of Applied Probability*, 20(6):2204–2260, 2010.

[43] W. Kang and K. Ramanan. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Annals of Applied Probability*, 22(2):477–521, 2012.

[44] H. Kaspi and K. Ramanan. Law of large numbers limits for many-server queues. *Annals of Applied Probability*, 21(1):33–114, 2011.

[45] J. Kim, R. S. Randhawa, and A. R. Ward. Dynamic scheduling in a many-server multi-class system: The role of customer impatience in large systems. *Manufacturing & Service Operations Management*, 20(2):285–301, 2018.

[46] J. S. H. Van Leeuwaarden, B. W. J. Mathijsen, and B. Zwart. Economies-of-scale in resource sharing systems: Tutorial and partial review of the QED heavy-traffic regime, 2017. arXiv:1706.05397v1.

[47] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. IEEE *Transactions on Networking*, 21:1378–1391, 2013.

[48] Y. Liu. Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Operations Research*, 66(2):514–534, 2018.

[49] Y. Liu and W. Whitt. The $\mathrm{G}_t/\mathrm{GI}/s_t+\mathrm{GI}$ many-server fluid queue. *Queueing Systems*, 71(4):405–444, 2012.

[50] Y. Liu and W. Whitt. Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research*, 256(2):473–486, 2017.

[51] Z. Long and J. Zhang. Convergence to equilibrium states for fluid models of many-server queues with abandonment. *Operations Research Letters*, 42(6-7):388–393, 2014.

[52] A. Mandelbaum and P. Momcilovic. Queues with many servers and impatient customers. *Mathematics of Operations Research*, 37(1):41–65, 2012.

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

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

[55] A. Mandelbaum and S. Zeltyn. Data stories about (im)patient customers in tele-queues. *Queueing Systems*, 73(4):1–32, 2013.

[56] J. A. Van Mieghem. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *The Annals of Applied Probability*, pages 809–833, 1995.

[57] D. Mukherjee, Y. Li, and D. A. Goldberg. Large deviations analysis for the $\mathrm{M/H_2}/n+\mathrm{M}$ queue in the Halfin-Whitt regime, 2018. arXiv:1803.01082v1.

[58] M. L. Pinedo. *Scheduling: theory, algorithms, and systems*. Springer Science & Business Media, 2012.

[59] A. Puha and A. R. Ward. A fluid limit for a multi-class many-server queue with general reneging distribution, 2018. Working Paper, To be posted on arXiv soon.

[60] J. E. Reed and T. Tezcan. Hazard rate scaling for the $\mathrm{GI/M}/n+\mathrm{GI}$ queue. *Operations Research*, 60(4):981–995, 2012.

[61] S. M. Ross. *Introduction to Probability Models*. Elsevier, 2010. 10th Edition.

[62] W. Smith. Various optimizers for single-stage production. *Naval Research Logistics Quarterly*, 3:59–66, 1956.

[63] A. R. Ward. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science*, 16(1):1–14, 2011.

[64] W. Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54(1):37–54, 2006.

[65] A. Wu, A. Bassamboo, and O. Perry. Service systems with dependent service and patience times. *Management Science*, 65(3):1151–1172, 2019.

[66] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems*, 51(3/4):361–402, 2005.

[67] J. Zhang. Fluid models of many-server queues with abandonment. *Queueing Systems*, 73(2):147–193, 2013.

[68] A. W. Zuñiga. Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions. *Stochastic Processes and their Applications*, 124(3):1436–1468, 2014.