Chapter ??

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

Amber 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 that fluid model to approximately solve scheduling control problems when the arrival rate remains constant over time. 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, that is consistent with the simple index rule known as $c\mu$, when the objective is to minimize the long-run average abandonment rate. 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 [56], 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 customer are impatient, and will abandon the system without receiving service if their wait becomes too long.

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 time 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 “$N \in \{1, 2,\ldots\}$” 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-in-first-out (FIFO), 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.

![Figure 1](image_url) 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. (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 FIFO and randomizing over which class receives service by flipping an unbiased coin when customers from both classes are waiting, result in both classes experiencing similar amounts of congestion. These observations are true regardless of the patience time 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 time distribution. In contrast, Figures 2 (a) and (b) show that the average number of customer abandonments is not affected by the patience time distribution, an observation supported by the rough calculation that when priority is given to one class, the other class’s abandonment percentage should be approximately

$$\frac{\text{low priority class arrival rate} - \text{remaining service capacity for that class}}{\text{low priority class arrival rate}} = \frac{60 - 40}{60} = \frac{1}{3}.$$
Figures 1 and 2 lead us to wonder how much the patience time distribution should affect the scheduling. The figures suggest that when the system manager only cares about abandonments, the scheduling should give priority to the most important class(es). More generally, 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 other scheduling polices must be considered. 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 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 presented here arises in [57] 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 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 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 $1/\theta_j < \infty$ to find his patience time and from the distribution determined by cdf $G^s_j$ having mean $1/\mu_j < \infty$ 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.](image)

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 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. 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 $a_j, j \in \mathbb{J}$. The process $R_j(T, \pi)$ tracks the cumulative number of abandonments from class $j \in \mathbb{J}$ under scheduling policy $\pi$.

Our objective is to find a scheduling policy $\pi$ that minimizes the long-run average cost:

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

The class of HL scheduling policies $\Pi$ 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
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 \( G/GI/N \) queue state space, which we do next.

**Assumption 1.** We assume \( E_j \) is a counting process (that is, a non-negative, non-decreasing 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 \( G_j^r \) and \( G_j^s \) are absolutely continuous with density functions \( g_j^r \) and \( g_j^s \) that have (possibly infinite) right edges of support:

\[
H_j^s := \sup\{x \in [0, \infty) : 1 - G_j^s(x) > 0\} \quad \text{and} \quad H_j^r := \sup\{x \in [0, \infty) : 1 - G_j^r(x) > 0\}, \quad \text{for each} \quad j \in \mathbb{J}.
\]

We let

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

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

**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 \( H \in \mathbb{R}_+ \), \( \mathbf{M}[0, H] \) 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 in 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 customer in service has been in service, and is also known as the age-in-service;
- \( \eta_j(t) \in \mathbf{M}[0, H_j^r) \), shown in Figure 4(b), stores the amount of time that has passed between each customer’s arrival time up until that customer’s potential abandonment time (that 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.

The fourth bullet point requires additional explanation. The measure \( \eta_j, j \in \mathbb{J} \), 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 [57]. 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
(a) The age-in-service measure $\nu_j$.
(b) The potential queue measure $\eta_j$.

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

$(\alpha(t), X(t), \nu(t), \eta(t))$. Then, the number of class $j \in \mathbb{J}$ customers in service at time $t \geq 0$ is found by integrating the measure $\nu_j(t)$ 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

$$\int_0^{H_j^s} \nu_j(t)(dx).$$

From the above display, the number of class $j \in \mathbb{J}$ customers waiting in queue is

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

Similar reasoning as above shows that the number of class $j$ customers whose potential abandonment times have not yet passed 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, and so is an upper bound on the number of class $j$ customers actually waiting in the queue; i.e.,

$$X_j(t) - \int_0^{H_j^s} \nu_j(t)(dx) \leq \int_0^{H_j^r} \eta_j(t)(dx).$$

3. The Fluid Model

The input to the fluid model is an arrival function $E$ having domain $\mathbb{R}_+$ and range $\mathbb{R}_+$. The arrival function is continuous, and each component has initial value zero and is non-decreasing. The arrival function arises as the fluid limit in 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

$$\mathbb{P} \left( \max_{j \in \mathbb{J}} \lim_{N \to \infty} \sup_{0 \leq t \leq T} \left| \frac{E_j^N(t)}{N} - E_j(t) \right| \right) = 1, \text{ for any } T \in (0, \infty).$$
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 $E_j(t) = \lambda_j t$, 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 non-negative 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 $\Pi$; see Theorem 4.1 in [57].

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 preceding paragraph, $E_j^N$ denoted the class $j$ arrival process to the $G/GI/N + GI$ queue, and $E_j$ denotes the $j$th component of the fluid model arrival function $E$. 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;
- $\nu_j(t) \in \mathcal{M}[0, H_j^+)$ approximates the scaled age-in-service measure;
- $\eta_j(t) \in \mathcal{M}[0, H_j^+)$ approximates the scaled potential queue measure.

We set $X := \mathbb{R}_+^J \times (\times_{j=1}^J \mathcal{M}[0, H_j^+)) \times (\times_{j=1}^J \mathcal{M}[0, H_j^+))$, and endow $X$ with the product topology, which is a Polish space (recalling that $\mathcal{M}[0, H_j^+)$ and $\mathcal{M}[0, H_j^+)$, 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 $\mathbf{C}(X)$, which denotes the set of functions having domain $\mathbb{R}_+$ and range $X$ that are continuous in time. Any fluid model solution is in $\mathbf{C}(X)$.

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 integrable with respect to the measure $\xi \in \mathcal{M}[0, H]$, 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 \( \xi \) is the exponential distribution with rate parameter \( \theta \), 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 \((X, \nu, \eta) \in C(\mathbb{X})\) for \( E \) satisfies the finiteness conditions

\[
\int_0^t \langle h_j^s, \nu_j(u) \rangle du < \infty \quad \text{and} \quad \int_0^t \langle h_j^r, \eta_j(u) \rangle du < \infty,
\]

has an initial potential queue measure with no atoms

\[
\langle 1_{\{x\}}, \eta_j(0) \rangle = 0 \quad \text{for all} \quad x \in [0, H_j^r),
\]

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;
- \( R_j(t) \) 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 is determined by the scheduling policy.

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

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

\[
B_j(t) := \langle 1, \nu_j(t) \rangle = \int_0^{H_j^r} \nu_j(t)(dx), \quad \text{for} \quad j \in \mathbb{J} \quad \text{and} \quad t \in \mathbb{R}_+.
\]

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

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

which we require to be non-negative 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

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

The measure \( \eta_j(t) \) is used to track the cumulative amount of fluid that abandons without ever having entered service. To see how to do this, first recall from the explanation of the potential queue measure, that the measure \( \eta_j \) 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 \( \chi_j \) given by

\[
\chi_j(t) := \inf \{ x \in \mathbb{R}_+ : \langle 1_{[0,x]}, \eta_j(t) \rangle \geq Q_j(t) \}, \quad \text{for} \quad j \in \mathbb{J} \quad \text{and all} \quad t \in \mathbb{R}_+,
\]
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 \( j \in \mathcal{J} \), which implies

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

see Remark 1 below. For any \( w \in [0, \chi_j(t)] \), the value of the hazard rate \( h_j^*(w) \) dictates the rate at which fluid with age \( w \) abandons in the next instant. This leads to cumulative abandonment

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

(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

\[
R_j(t) = \int_0^t \theta_j \eta_j(u)[0, \chi_j(u)] du = \int_0^t \theta_j Q_j(u) du, \text{ for } j \in \mathcal{J} \text{ and all } t \in \mathbb{R}_+.
\]

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

\[
\langle h_j^*, \nu_j(u) \rangle = \int_0^{H_j^*} h_j^*(x) \nu_j(u)(dx)
\]

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

\[
D_j(t) := \int_0^t \langle h_j^*, \nu_j(u) \rangle du, \text{ for } j \in \mathcal{J} \text{ and all } t \in \mathbb{R}_+.
\]

(11)

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

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

(12)

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

\[
X_j(t) = X_j(0) + E_j(t) - R_j(t) - D_j(t), \text{ for } j \in \mathcal{J} \text{ and } t \in \mathbb{R}_+.
\]

(13)

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

\[
\langle f, \nu_j(t) \rangle = \left( f(\cdot + t) \frac{1 - G^*_j(\cdot + t)}{1 - G^*_j(\cdot)} , \nu_j(0) \right) + \int_0^t f(t-u)(1 - G^*_j(t-u)) dK_j(u),
\]

(14)

\[
\langle f, \eta_j(t) \rangle = \left( f(\cdot + t) \frac{1 - G^*_j(\cdot + t)}{1 - G^*_j(\cdot)} , \eta_j(0) \right) + \int_0^t f(t-u)(1 - G^*_j(t-u)) dE_j(u).
\]

(15)
Definition 1. Given an arrival function $E$, a fluid model solution for $E$ is $(X, \nu, \eta) \in C(X)$ that satisfies (5) and (6), has auxiliary functions defined by (7)-(12) that satisfy $B, Q, R, D, K \in C(R^+_J)$, 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 $t \in R_+$;
(b) The function $Q$ satisfies $0 \leq Q_j(t) \leq (1, \eta_j(t))$ for all $j \in J$ and $t \in R_+$, 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 $j \in J$;
(d) The measure-valued functions $\nu$ and $\eta$ 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 service entry function $K$, that would connect to a defined scheduling policy in the class II 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 service entry 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 [57]. 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 $t \in R_+$; see [57, Lemma 3.4]. Then, for all $t \geq 0$ and $j \in \mathbb{J}$,

$$R_j(t) = \int_0^t \theta_j \eta_j(u) \, 0, \chi_j(u)) \, du = \int_0^t \theta_j \eta_j(u) \, 0, \chi_j(u)) \, du.$$  

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

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 equation $\max(X_1(t) - 1, 0) = Q_1(t)$ for all $t \in R_+$. 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 [66, Theorems 3.5 and 4.4] for an extension that does not require the absolute continuity of the service and patience time distributions, [65, Theorems 3.1 and 3.3] for an alternative approach 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 service entry function $K$ 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 service entry function $K$, meaning there is no expectation of uniqueness.
Remark 4. (A No Abandonment Model) For a multiclass $G/GI/N$ queue with no abandonments, the relevant fluid model eliminates the measure $\eta$, and has state space $\mathcal{X} := \mathbb{R}_+^J \times (\times_{j=1}^J \mathbf{M}[0, H_j^+] )$. Then, Definition 1 is modified as follows: Given an arrival function $E$, a fluid model solution for $E$ is $(X, \nu) \in \mathbb{C}(\mathcal{X})$ that satisfies $\int_0^t \langle h_j^*, \nu_j(u) \rangle \, du < \infty$, has auxiliary functions defined by (7), (8), (11), and (12) that satisfy $B, D, Q, K \in \mathbb{C}(\mathbb{R}_+^J)$, 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 $j \in \mathcal{J}$ and $t \in \mathbb{R}_+$.

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 $G/GI/N$ queue are independently sampled from a given distribution. Then, the arrival processes are renewal, and the queue is a $GI/GI/N + GI$ queue. Recall that in this case the condition (4) holds when the renewal arrival processes have rate $\lambda_j N$ for each $j \in \mathcal{J}$. 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

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

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. 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 $(X^*, \nu^*, \eta^*) \in \mathcal{X}$ is said to be an invariant state for the arrival function (16) if the constant function $(X(t), \nu(t), \eta(t)) = (X^*, \nu^*, \eta^*)$ 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 $\lambda$. 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 time 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 [59]) and is given by

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

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

An invariant state is uniquely specified by the allocation of server capacity given to each class. The server effort allocation to class $j$ should not exceed the workload contribution from class $j$,

$$\rho_j := \lambda_j / \mu_j, \text{ for } j \in \mathcal{J}.$$

Note that a forthcoming proof correction may require some additional assumptions not present in the original manuscript.
In other words, the server effort allocations should lie in the set
\[ B := \left\{ b \in \mathbb{R}_+^J : \text{for all } j \in J \text{ and } \sum_{j=1}^{J} b_j \leq 1 \right\}. \]

**Theorem 1.** (Fluid Model Invariant States) Suppose that \( G_j^* \) is strictly increasing for each \( j \in J \), with inverse function \( (G_j^*)^{-1} \). For \( b \in B \), define
\[
q_j(b_j) := \begin{cases} 
\frac{\lambda_j}{b_j}, & \text{if } b_j = 0, \\
\frac{\lambda_j}{b_j} G_{e,j}^* \left( (G_j^*)^{-1} \left( 1 - \frac{b_j}{\mu_j} \right) \right), & \text{if } b_j \in (0, \rho_j] 
\end{cases} \quad \text{for } j \in J, \tag{17}
\]
and, for each \( j \in J \), let
(i) \( \eta_j^*(dx) := \lambda_j \left( 1 - G_j^*(x) \right) dx \) for each \( x \in \mathbb{R}_+ \),
(ii) \( \upsilon_j^*(dx) := b_j \mu_j \left( 1 - G_j^*(x) \right) dx \) for each \( x \in \mathbb{R}_+ \),
(iii) \( X_j^* := b_j + q_j(b_j) \).

Then \( (X^*, \upsilon^*, \eta^*) \in I_\Lambda, \ B^* = b, \) and \( Q^* = q \). Conversely, if \( (X^*, \upsilon^*, \eta^*) \in I_\Lambda, \) then \( B^* \in B \) and \( (X^*, \upsilon^*, \eta^*) \) 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 \([63, \text{Equation (3.7)}]\).

**Remark 5.** (Related Results) A version of Theorem 1 is proved in \([43, \text{Theorem 5.5}]\) and in \([65, \text{Theorem 3.2}]\) when \( J = 1 \), meaning the optimal scheduling policy is trivial, as observed in Remark 2. A version of Theorem 1 is also proved in \([9, \text{Theorem 3.3}]\) when the fluid model has added equations relevant for a static priority scheduling policy. However, not all \( b \in B \) can be achieved. This motivates the need to understand the fluid limits associated with a wider range of scheduling policies.

**Remark 6.** (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 \([50, \text{Theorems 1 and 2}]\), but requires somewhat restrictive conditions on the initial state when there is not enough capacity to serve all fluid; see also \([43, \text{Section 7.1}]\) for more discussion of the relevant issues. Proving such a convergence result for \( J > 1 \) is an open problem.

**Remark 7.** (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 \( \lambda_j \), as shown in (16). This results in the predicted class \( j \in J \) 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 \([57]\), the service rate \( \mu_j \) and mean patience times \( 1/\theta_j \) appearing in the formula for \( q_j \) are not scaled, \( j \in J \).

Figure 5 shows the fluid queue sizes are ordered by the variance of the patience time distributions. Figure 5(a) assumes the patience time distribution is a mean 1 Gamma distribution having both shape and rate parameters equaling \( p \in \{0.2, 0.5, 1, 2, 5\} \). The variance of a Gamma(\( p \)) distribution is \( 1/p \), which is decreasing in \( p \). Figure 5(b) assumes the patience time 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

\[ G(x) = \frac{(p x e^{-x})^{p-1} e^{-x}}{\Gamma(p) p^{p-1} e^{-p}} dt \]

\[ \int_0^t e^{-x} dx \]
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 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 leads to 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 \(\lambda_j\) 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

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

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

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

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,

\[
R^*_j(T) = T \int_0^{\chi^*_j} h_j^*(w) \eta^*_j(dw) \text{ for } j \in J \text{ and all } T \in \mathbb{R}^+. \tag{21}
\]

We derive the expression for \(\chi^*_j\) for any \(j \in J\) in this paragraph, which is necessary to simplify (21). For this, from (9) and the fact that \(\eta^*_j\) has no atoms (recall Remark 1), \(\chi^*_j\) solves

\[
\left\{1_{[0, \chi^*_j]}, \eta^*_j\right\} = Q^*_j.
\]
Equivalently, substituting the definitions of $\eta^*_j$ and $Q^*_j$ given in Theorem 1, we have

$$\int_0^{\chi^*_j} \lambda_j (1 - G^*_j(y)) \, dy = \frac{\lambda_j}{\theta_j} G^*_e,j \left( (G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right) \right).$$

Substituting for the excess life distribution yields

$$G^*_e,j \left( (G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right) \right) = \int_0^{(G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right)} \theta_j (1 - G^*_j(y)) \, dy.$$

The previous two displays imply $\chi^*_j$ satisfies

$$\int_0^{\chi^*_j} (1 - G^*_j(y)) \, dy = \int_0^{(G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right)} (1 - G^*_j(y)) \, dy,$$

from which we find

$$\chi^*_j = (G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right). \tag{22}$$

Finally, to see (19), for any $j \in J$, we substitute $\chi^*_j$ in (22) and $\eta^*_j$ given in Theorem 1 into (21) to find

$$R^*_j(T) = T \int_0^{(G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right)} h^*_j(y) \lambda_j (1 - G^*_j(y)) \, dy.$$

The definition of the hazard rate and straightforward calculation show that

$$R^*_j(T) = T \int_0^{(G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right)} \lambda_j g^*_j(y) \, dy$$

$$= T \lambda_j G^*_j(y) \bigg|_{y=0}^{(G^*_j)^{-1} \left( 1 - \frac{b_j}{\rho_j} \right)}$$

$$= T \lambda_j \left( 1 - \frac{b_j}{\rho_j} \right).$$

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

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

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 0.
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 \( \{j^*+1, \ldots, J\} \) are not served at all, in an asymptotic sense.

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

- \( q_j (b^*_j) = 0 \) for \( j \in \{1, \ldots, j^*-1\} \);
- \( q_{j^*} (b^*_j) = \left( 0, \frac{\lambda_{j^*}}{\theta_{j^*}}, G^c_{e,j^*}(H^r_{j^*}) \right) \);
- \( q_j (b^*_j) = \lambda_j \theta_j \) 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 \( M/M/1 + M \) queue. In particular, if the patience time distribution of class \( j \in \mathbb{J} \) is exponential, then \( q_{j^*} (b^*_j) = 0 \) for \( j \in \{1, \ldots, j^*-1\} \), and

\[ q_{j^*} (b^*_j) = \frac{\lambda_{j^*} - b^*_j \mu_{j^*}}{\theta_{j^*}} \quad \text{and} \quad q_j (b^*_j) = \frac{\lambda_j}{\theta_j} \quad \text{for } j \in \{j^*+1, \ldots, J\}. \number(25)\]

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 time 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 time distribution is 1/\( \theta_j \)). In that case,

\[ G^c_{e,j}(x) = \frac{\theta_j}{2} x, \quad \text{for } x \in \left[ 0, \frac{2}{\theta_j} \right] \quad \text{and} \quad (G^c_{e,j})^{-1}(x) = \frac{2}{\theta_j} x, \quad \text{for } x \in [0, 1]. \]

The associated excess life distribution is

\[ G^e_{e,j}(x) = \int_{0}^{x} \theta_j \left( 1 - \frac{\theta_j}{2} y \right) dy = \theta_j x \left( 1 - \frac{\theta_j}{4} x \right), \quad \text{for } x \in \left[ 0, \frac{\theta_j}{2} \right]. \]
We then calculate 
\[ G_{r,j} \left( \left( (G_j^r)^{-1} (x) \right) \right) = 2x - x^2, \text{ for } x \in [0,1], \]
which shows 
\[ q_j^*(b_j^*) = \frac{\lambda_j}{\theta_j} \left[ 2 \left( 1 - \frac{b_j^*}{\rho_j} \right) - \left( 1 - \frac{b_j^*}{\rho_j} \right)^2 \right] = \frac{\lambda_j}{\theta_j} \left( 1 - \left( \frac{b_j^*}{\rho_j} \right)^2 \right) \]  
(26)

and 
\[ q_j(b_j^*) = \frac{\lambda_j}{\theta_j} \text{ for } j \in \{j^* + 1, \ldots, J\}. \]

**Interpreting the Fluid Control Problem Solution as a Scheduling Policy**

The fluid solution \( b^* \) that solves (23) suggests that a static priority scheduling policy \( \pi_{sp} \) 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.

![Graphs showing abandonment rates](image)

**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(1,4) patience time 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 \( \pi_{sp} \) does not depend on the patience time 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 time 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 \( \pi_{sp} \). Figure 6\(^3\) shows numerically that neither the patience time distribution nor the service time distribution have much impact on the rate at which customers abandon. In Figure 6, the two service distributions considered are Lognormal(1,4) and Gamma(2)\(^4\) defined at the end of Section 4, which both have mean 1, but have variance 4 and 0.5 respectively, and the patience time distributions are those considered in Figure 5.

---

\(^3\) 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 are the average over 10 runs, and the error bars shows the 95% confidence intervals over the 10 runs.

\(^4\) The service time distributions we assume in Figure 6 are the ones used in Table 1 in [63] in the single class setting, to illustrate the lack of effect of the service distribution on steady-state system performance measures for overloaded systems.
The class 2 mean steady-state queue-length does depend on the patience time 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 \( j \in \mathcal{J} \) provides a good approximation of the class 2 queue-lengths. Specifically, for the system in Table 1, if we recall Remark 7 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 time 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

\[
\mathbb{E}[Q_1^N] \approx Nq_1(0.6) = 0 \quad \text{and} \quad \mathbb{E}[Q_2^N] \approx Nq_2(0.4) = 0.6G_{r,j}^{-1} \left( (G_j^r)^{-1} \left( 1 - \frac{0.4}{0.6} \right) \right).
\]

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Approximated</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v )</td>
<td>( \mathbb{E}[Q_1^N] )</td>
</tr>
<tr>
<td>0.2</td>
<td>42.1120</td>
</tr>
<tr>
<td>0.5</td>
<td>33.3982</td>
</tr>
<tr>
<td>1</td>
<td>25.9555</td>
</tr>
<tr>
<td>2</td>
<td>18.8790</td>
</tr>
<tr>
<td>5</td>
<td>11.4800</td>
</tr>
</tbody>
</table>

Table 1. A comparison of the queue-length approximations given in Theorem 4.2 with their simulated steady-state mean in a \( M/\text{LN}(1,4)/100 + \text{LN}(1, v) \) 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 time distribution increases, which is consistent with the theoretic prediction in Figure 5. This effect becomes even more pronounced when the patience time 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 system size becomes larger, consistent with the underlying theory in [57], we do not know the connection between the rate at which the prediction error becomes small and the patience time distribution characteristics.

**Remark 8.** (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 [60] 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 \( Q_j(t, \pi) \), \( j \in \mathbb{J} \) 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

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

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

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

Define the modified cost that incorporates both holding and abandonment

\[ \tilde{c}_j = a_j + \frac{c_j}{\theta_j}, \]
and observe that (28) becomes
\[
\inf_{b \in \mathbb{R}} \hat{c}_j (\lambda_j - b_j \mu_j). \tag{29}
\]
In particular, the solution to the LP (28) is exactly as in Section 6, except with \(\hat{c}_j\) replacing \(c_j\) for \(j \in \mathbb{J}\), as observed in [6]. This suggests that the static priority \(\hat{c}_j \mu\) scheduling policy that ranks classes in the order of their \(\hat{c}_j \mu_j\) 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 time 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 strictly increasing and continuously differentiable on \((0, 1)\) for all \(j \in \mathbb{J}\). If \(h_j^r\) is (strictly) increasing, 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 9. (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 \(C_N(\pi)\) is the long-run average cost when the number of servers is \(N\), we would like to show that the class of scheduling policies \(\Pi\) is such that
\[
\lim_{N \to \infty} \frac{C_N(\pi)}{N} \geq m^*, \quad \text{for all } \pi \in \Pi, \tag{30}
\]
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 times all follow exponential distributions), (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 times are 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. 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

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

(32)

recalling $\lambda_j^N$ 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 $E_j^N, j \in J,$ 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 ($J = 1$) and large arrival rate, 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. Furthermore, HL scheduling (equivalently, first-come-first-serve) minimizes the fluid queue-lengths [12, Corollary 1]. However, caution is warranted, because 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], and 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].

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 customer patience times do not follow an exponential distribution. 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.

**Time-Varying Arrival Rates**

The fluid model in Section 3 allows for time-varying customer arrival rates. Moreover, the fluid model can be modified to many server systems in which customers do not abandon, as detailed in Remark 4 (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 [46, 25], call centers [1, 26], and hospital operations [16, 37, 24, 17], all of which experience time-varying demand.

---

5 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 3.9], and an erratum is currently in progress.
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, 60], continuing with the later work of [54] (which provides an excellent literature review), [52], 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, 47, 49], and related papers discuss how to calculate fluid performance measures [48], 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 $N$, $\lambda_j$, and $\mu_j$ 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 $\beta$ 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 $\beta$ 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 $\beta$ 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 time distributions are exponential in [27], and to non-exponential distributions in [20, 23, 51, 58, 64]; for a comprehensive review of the QED regime see the tutorial papers [21, 61], for very recent work in the QED regime see the large deviations analysis in [55], and for an overview of the different regimes see the survey papers [62, 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 [53] 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 inter-arrival, 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

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

Then, by (17),

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

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 \( (X, \nu, \eta) \) be the constant function such that \( (X(t), \nu(t), \eta(t)) = (X^*, \nu^*, \eta^*) \) for all \( t \geq 0 \). We must show that \( (X, \nu, \eta) \) is a fluid model solution; i.e., that \( (X, \nu, \eta) \in \mathbb{C}(\mathbb{X}) \), that \( B, Q, R, D, K \in \mathbb{C}(\mathbb{R}_+^4) \), and that (5), (6), and (a)-(d) in Definition 1 hold. First observe that (i) and (ii) imply that

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

and

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

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 \( t \geq 0 \),

\[ B_j(t) = B^*_j = b_j, \quad \text{by (ii) and (7)}, \]

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

\[ D_j(t) = b_j \mu_j t, \quad \text{by (ii) and (11)}, \]

\[ K_j(t) = b_j \mu_j t, \quad \text{by (12)}. \]

From (33)-(36), \( B, Q, R, D, K \in \mathbb{C}(\mathbb{R}_+^4) \), and from (i)-(iii) \( (X, \nu, \eta) \in \mathbb{C}(\mathbb{X}) \). The fact that \( R \in \mathbb{C}(\mathbb{R}_+^4) \) 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 \( 0 \leq Q_j(t) \leq (1, \eta_j) \) for all \( j \in \mathbb{J} \) and \( t \in \mathbb{R}_+ \) follows from (34), because \( f_j \) has range \([0,1]\) and \( (1, \eta_j) = \lambda_j/\theta_j \) for all \( j \in \mathbb{J} \) from (i). Next, for each \( j \in \mathbb{J} \), let \( \chi_j \) be the unique solution of

\[ G^*_{e,j}(\chi_j) = f_j(1 - b_j/\rho_j) \quad \text{if } b_j > 0 \text{ and } H^*_j \text{ otherwise}. \]
Then, letting \((G_{\nu,j}^x)^{-1}\) be the inverse function for \(G_{\nu,j}^x\),
\[
\chi_j = (G_{\nu,j}^x)^{-1}(f_j(1-b_j/\rho_j)) = (G_j^y)^{-1}(1-b_j/\rho_j),
\]
for each \(j \in J\) such that \(b_j > 0\),
and so by (i), (9), and (10),
\[
R_j(t) = \lambda_j \int_0^t \int_0^{x_j} g_j^y(x)dx
\]
\[
= \begin{cases} 
\lambda_j t, & \text{if } b_j = 0, \\
\lambda_j G_j^y(\chi_j)t, & \text{if } b_j \in (0, \rho_j], \\
(\lambda_j - b_j \mu_j)t, & \text{if } b_j \in (0, \rho_j],
\end{cases}
\]
for each \(j \in J\) and \(t \geq 0\). By (37) and the expression for \(D_j\) in (35), \(D_j(t) + R_j(t) = \lambda_j t = E_j(t)\) for all \(t \geq 0\) and \(j \in J\), and so \((X, \nu, \eta)\) 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 \(j \in J\), and \(t \geq 0\),
\[
\int_0^\infty \frac{1 - G_j^x(x + t)}{1 - G_j^x(x)} f(x + t)\nu_j^*(dx) + \int_0^t (1 - G_j^x(t - u))f(t - u)dK_j(u)
\]
\[
= b_j \mu_j \int_0^\infty (1 - G_j^x(x + t)) f(x + t)dx + b_j \mu_j \int_0^t (1 - G_j^x(u))f(u)du
\]
\[
= b_j \mu_j \int_0^\infty (1 - G_j^x)(u)du
\]
\[
= \langle f, \nu_j^* \rangle.
\]
Since \(\nu_j(t) = \nu_j^*\) for all \(t \geq 0\) and \(j \in J\), (14) holds. Finally, by (i), for any continuous and bounded function \(f\) having domain \(\mathbb{R}_+\), for all \(j \in J\), and \(t \geq 0\),
\[
\int_0^\infty \frac{1 - G_j^x(x + t)}{1 - G_j^x(x)} f(x + t)\eta_j^*(dx) + \lambda_j \int_0^t (1 - G_j^x(t - s))f(t - s)ds = \langle f, \eta_j^* \rangle.
\]
Since \(\eta_j(t) = \eta_j^*\) for all \(t \geq 0\) and \(j \in J\), (15) holds.

In summary, we have established \((X, \nu, \eta) \in C(\mathbb{R})\) satisfies (5) and (6), and that \(B, Q, R, D, K \in \mathbb{C}(\mathbb{R}_+^J)\) 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 \((X, \nu, \eta)\) be the constant function such that \((X(t), \nu(t), \eta(t)) = (X^*, \nu^*, \eta^*)\) for all \(t \geq 0\). By assumption \((X, \nu, \eta)\) is a fluid model solution. Therefore, from Definition 1(a), \(0 \leq B_j^* \leq 1\). In order to show that \(B^* \in \mathbb{B}\), we must also show that \(B_j^* \leq \rho_j\) for each \(j \in J\). We will verify this in the process of verifying that \((X^*, \nu^*, \eta^*)\) 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 \(\mathbb{R}_+, j \in J, t \geq 0\),
\[
\langle f, \eta_j^* \rangle = \int_0^\infty \frac{1 - G_j^x(x + t)}{1 - G_j^x(x)} f(x + t)\eta_j^*(dx) + \lambda_j \int_0^t (1 - G_j^y(t - u))f(t - u)du.
\]
For each \(j \in J\), as \(t \to \infty\), the first integral converges to zero by dominated convergence, and the second to \(\lambda_j \int_0^\infty (1 - G_j^x(u))f(u)du\). 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, \( 0 \leq Q^*_j \leq \langle 1, \eta^*_j \rangle = \lambda_j/\theta_j \) for each \( j \in J \). Define \( \chi_j \in [0, H^*_j] \) to be the unique solution to

\[
Q^*_j = \int_0^{\chi_j} \eta^*_j(u)du, \quad \text{for each } j \in J.
\]

so that \( \chi_j \) satisfies (9). We fix \( j \in J \) and separate the cases that \( \chi_j = H^*_j \) and \( \chi_j < H^*_j \).

Case \( \chi_j = H^*_j \). From (i) and (10),

\[
R_j(t) = \int_0^t \int_0^{H^*_j} h^*_j(w)\lambda_j(1 - G^*_j(w))dw = \lambda_j t, \quad \text{and all } t \in \mathbb{R}_+.
\]

From (13),

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

and from (12),

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

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

\[
\langle f, \nu^*_j \rangle = \int_0^{H^*_j} \frac{1 - G^*_j(x + t)}{1 - G^*_j(x)} f(x + t)\nu^*_j(dx) \to 0 \text{ as } t \to \infty, \quad \text{for all } t \in \mathbb{R}_+, \tag{38}
\]

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 \( \chi_j < H^*_j \). From (i) and (9), \( Q^*_j = \lambda_j G^*_{e,j}(\chi_j)/\theta_j \), or, equivalently, \( \chi_j = (G^*_{e,j})^{-1}(\theta_j Q^*_j/\lambda_j) \). From (i) and (10),

\[
R_j(t) = \int_0^t \left( \int_0^{\chi_j} h^*_j(w)\lambda_j(1 - G^*_j(w))dw \right) du = \lambda_j G^*_j \left( (G^*_{e,j})^{-1}(\theta_j Q^*_j/\lambda_j) \right) t,
\]

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

\[
k^*_j = \lambda_j \left( 1 - G^*_j \left( (G^*_{e,j})^{-1}(\theta_j Q^*_j/\lambda_j) \right) \right).
\]

From (13),

\[
D_j(t) = k^*_jt, \quad \text{for all } t \in \mathbb{R}_+,
\]

and from (12),

\[
K_j(t) = k^*_jt, \quad \text{for all } t \in \mathbb{R}_+.
\]

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

\[
\langle f, \nu^*_j \rangle = \int_0^\infty \frac{1 - G^*_j(x + t)}{1 - G^*_j(x)} f(x + t)\nu^*_j(dx) + k^*_j \int_0^t (1 - G^*_j(t - u))f(t - u)du, \tag{39}
\]

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 \( k^*_j \int_0^\infty (1 - G^*_j(u))f(u)du \). Hence

\[
\nu^*_j(du) = k^*_j(1 - G^*_j(u))du \text{ for all } u \in \mathbb{R}_+.
\]

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 \( j \in J \) such that case (i) holds, \( B^*_j = 0 \), and for any \( j \in J \) such that case (ii) holds, \( k^*_j \leq \lambda_j \) implies \( B^*_j \leq \rho_j \).
Finally, to see (iii) of Theorem 1 holds, observe that from the argument used to establish (ii) above, for any \( j \in J \), if \( \chi_j = H_j^* \), then \( B_j^* = 0 \) and
\[
Q_j^* = \lambda_j / \theta_j.
\]
Otherwise, if \( \chi_j < H_j^* \), then the definition of \( k_j^* \) and the fact that \( B_j^* = k_j^* / \mu_j \) implies
\[
B_j^* = \frac{\lambda_j}{\mu_j} \left( 1 - G_j^* \left( (G_{e,j}^*)^{-1} \left( \frac{\theta_j Q_j^*}{\lambda_j} \right) \right) \right).
\]
Re-arranging terms in the above equality shows
\[
Q_j^* = \frac{\lambda_j}{\theta_j} G_{e,j}^* \left( (G_j^*)^{-1} \left( 1 - \frac{B_j^*}{\rho_j} \right) \right)
\]
We conclude \( Q_j^* \) and \( B_j^* \) satisfy (17) for all \( j \in J \), which implies (iii) by (8).

9. Proof of Lemma 1

Fix \( j \in J \). By assumption, \( g_j^* \) is positive and continuous on \((0, H_j^*)\). In particular, \( G_j^* \) is continuously differentiable, it is almost everywhere differentiable on \((0, H_j^*)\) with positive derivative \( g_j^* \). Then, by the inverse function theorem, \((G_j^*)^{-1}\) is differentiable at each \( x \in (0, 1) \), and
\[
\frac{d}{dx} (G_j^*)^{-1}(x) = \frac{1}{g_j^* ((G_j^*)^{-1}(x))}.
\]
Also, by the fundamental theorem of calculus, \( G_{e,j}^* \) is continuously differentiable on \((0, H_j^*)\) with derivative
\[
\frac{d}{dx} G_{e,j}^*(x) = \theta_j (1 - G_j^*(x)), \text{ for } x \in (0, H_j^*).
\]
We use the chain rule in (17) to find that for \( b \in (0, \rho_j) \)
\[
q_j^*(b) = -\mu_j \left( 1 - G_j^* \left( (G_j^*)^{-1} \left( 1 - b/\rho_j \right) \right) \right) = \frac{-\mu_j}{h_j^* \left( (G_j^*)^{-1} \left( 1 - b/\rho_j \right) \right)}.
\]
As \( b \in (0, \rho_j) \) increases, \( 1 - b/\rho_j \) decreases, causing \((G_j^*)^{-1} \left( 1 - b/\rho_j \right) \) to decrease and in turn \( h_j^* \left( (G_j^*)^{-1} \left( 1 - b/\rho_j \right) \right) \) 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 \( h_j^* \) is strictly increasing, then \( q_j \) is strictly concave on \((0, \rho_j)\).

Acknowledgements

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


