Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Fluid Limits for Multiclass Many-Server Queues with General Reneging Distributions and Head-of-the-Line Scheduling

Amber L. Puha, Amy R. Ward

To cite this article:
Amber L. Puha, Amy R. Ward (2021) Fluid Limits for Multiclass Many-Server Queues with General Reneging Distributions and Head-of-the-Line Scheduling. Mathematics of Operations Research

Published online in Articles in Advance 21 Dec 2021

. https://doi.org/10.1287/moor.2021.1166

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

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–37
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Fluid Limits for Multiclass Many-Server Queues with General Reneging Distributions and Head-of-the-Line Scheduling

Amber L. Puha,[a] **Amy R. Ward**[b]

[a] Department of Mathematics, California State University San Marcos, San Marcos, California 92096; [b] The University of Chicago Booth School of Business, Chicago, Illinois 60637

Contact: apuha@csusm.edu, https://orcid.org/0000-0001-8988-3855 (ALP); amy.ward@chicagobooth.edu, https://orcid.org/0000-0003-2744-2960 (ARW)

Received: August 12, 2019
Revised: July 8, 2020
Accepted: March 23, 2021
Published Online in Articles in Advance: December 21, 2021

MSC2000 Subject Classification: Primary: 60K25, 60F17; secondary: 68M20, 90B22
OR/MS Subject Classification: Primary: Queues: limit theorems; secondary: probability: stochastic model applications

https://doi.org/10.1287/moor.2021.1166

Abstract. We describe a fluid model with time-varying input that approximates a multiclass many-server queue with general reneging distribution and multiple customer classes (specifically, the multiclass G/GI/N+GI queue). The system dynamics depend on the policy, which is a rule for determining when to serve a given customer class. The class of admissible control policies are those that are head-of-the-line (HL) and nonanticipating. For a sequence of many-server queues operating under admissible HL control policies and satisfying some mild asymptotic conditions, we establish a tightness result for the sequence of fluid scaled queue state descriptors and associated processes and show that limit points of such sequences are fluid model solutions almost surely. The tightness result together with the characterization of distributional limit points as fluid model solutions almost surely provides a foundation for the analysis of particular HL control policies of interest. We leverage these results to analyze a set of admissible HL control policies that we introduce, called weighted random buffer selection (WRBS), and an associated WRBS fluid model that allows multiple classes to be partially served in the fluid limit (which is in contrast to previously analyzed static priority policies).

Funding: Research was supported in part by the National Science Foundation [Grant DMS-1510198]. Financial support from the University of Chicago Booth School of Business is gratefully acknowledged.

Keywords: many-server queue • reneging • fluid limits • measure-valued process • scheduling control

1. Introduction

A classic question in the scheduling literature is to decide which customer should next go into service when a server becomes free (Pinedo [18]). The importance of answering that question is because scheduling has a nontrivial impact on the customer waiting times, which can result in impatient customers abandoning the system prior to entering service. One very appealing scheduling rule is static priority, because of its simplicity, easy implementation, and optimality properties. In static priority scheduling, customers within each class are served in a head-of-the-line (HL) fashion, with the customer within each class that has been waiting in system the longest being designated as the HL customer for that class. Classes are ranked and the HL customer next served is in accordance with that ranking, which is independent of the system state. The paper Atar et al. [3] shows that a static priority scheduling rule asymptotically minimizes the long-run average cost associated with holding and abandonment in an overloaded multiclass many-server queue with exponentially distributed interarrival, service, and patience times.

However, static priority scheduling may not be asymptotically optimal in general.[1] When the patience time distributions are not exponentially distributed, customers that have already waited some amount of time may be more or less willing to continue waiting. Static priority scheduling does not account for how customer willingness-to-wait changes over time, which suggests that a different class of scheduling rules may perform better, an observation that has been validated numerically in the single-server setting in Kim and Ward [13, figure 3] and in the many-server setting in Kim et al. [14, table 1]. The issue is that an asymptotically optimal rule may require partially serving multiple classes, which cannot happen under static priority. A static priority rule attempts to serve all its high-priority classes, ignores its low-priority classes, and partially serves at most one medium-priority class.

Another issue with static priority is that it can be perceived as "unfair." In particular, when there are multiple customers waiting, and a server becomes available, the server will next serve the highest-ranking HL customer, even if that customer only just arrived and other lower-ranking HL customers have been waiting for much

longer. This potential unfairness can prevent priority-based scheduling policies from being adopted, an issue discussed in Wierman [23] in the context of computer systems.

Fairness applies both to customers and to servers. Being fair to servers can mean allowing them to idle in the presence of waiting customers, in order to avoid overworking them. Moreover, when that desire to be fair to servers is captured in the objective function through a utilization cost, then an economically optimal limiting regime has intentional server idling; see Zhan and Ward [24, theorem 1 and example 1] for such a result in the single-class setting.

Intentional server idling can also be desirable when server fairness is ignored. For example, in the single-server setting in Afeche [1] and Afeche and Pavlin [2], when a firm jointly determines a price/lead-time menu and a scheduling policy in order to maximize revenue, intentionally idling servers can be necessary to exploit heterogeneous customer preferences. We conjecture a similar situation can arise in the many-server setting.

In summary, we are motivated to provide a framework for analyzing scheduling rules that have more flexibility than static priority in choosing how to serve classes. We would like that framework to be as general as possible. At a minimum, we would like that framework to include scheduling rules that (i) can partially serve multiple customers classes and (ii) may or may not be work conserving.

1.1. Contributions of This Paper

We develop a framework for analyzing the performance of a wide range of HL scheduling or control policies for the multiclass many-server queue with generally distributed interarrival, service, and patience times (i.e., a multiclass $G/GI/N{+}GI$ queue). We include the possibility that some customers will not abandon; that is, the patience time distributions may have atoms at infinity. We allow customers to arrive individually or in batches in a Markovian, possibly time-inhomogeneous fashion within each class.

1.1.1. Non-Policy-Specific Fluid Limits. We study the asymptotic behavior of the multiclass many-server queue for a large class of nonpreemptive HL control policies under fluid (functional law of large numbers) scaling. We refer to the class of HL control policies that we study as the admissible policy class, which are those policies that satisfy a collection of balance and evolution equations fundamental for multiclass many-server queues (see Definition 2), and under which the entry-into-service process is nonanticipating (see Definition 3).

For sequences of such systems under fluid scaling, we prove a tightness result that holds without the need to fully specify the particulars of the admissible HL control policy for each system (see Theorem 2), provided mild asymptotic conditions hold (see Assumptions 1–3). The value here is that, if one is interested to study a particular admissible HL control policy, then in order to obtain tightness under fluid scaling, they simply need to check that the aforementioned mild asymptotic conditions are satisfied.

To study fluid limits points, we formulate an associated fluid model, a set of fluid model equations that are fluid analogs of the evolution and balance equations that should be satisfied in the fluid limit by any given HL control policy. In particular, solutions to these fluid model equations are not unique because the control policy-specific dynamics are not accounted for in the fluid model equations. Under additional continuity assumptions, we show that distributional limit points of the aforementioned tight sequences are fluid model solutions almost surely (see Theorem 1). Because different convergent subsequences do not necessarily converge in distribution to a common limit, convergence of the entire sequence does not follow immediately from Theorems 1 and 2. Even so, Theorems 1 and 2 can be used as foundational results toward proving such convergence because the application of Theorem 1 implies that all fluid limit points satisfy the fluid model equations, including the highly nonlinear equation satisfied by the limiting reneging process (see (40)).

The specification of a particular admissible HL control policy that uniquely characterizes the system dynamics is required to prove convergence to a unique fluid limit. This requires adding one or more policy-specific equations and/or inequalities both to the multiclass many-server queue and to the fluid model. The additional work is to prove that the specification of the policy for the multiclass many-server queue (in the prelimit) does indeed give rise to the added fluid model equations in the fluid limit, and to prove uniqueness of fluid model solutions. In Section 4.4, a program detailing these steps and showing how to leverage the results in Theorems 1 and 2 to prove convergence for a specific admissible HL control policy is outlined and then illustrated with a well-studied example (static priority).

1.1.2. Analysis of the Weighted Random Buffer Selection Policy Class. In the final section of the paper, we introduce the weighted random buffer selection (WRBS) policy class and follow the program outlined in Section 4.4 to prove a weak convergence result. WRBS policies are easy to implement, and also satisfy the stated desires (i) and (ii) above, to include scheduling rules that can partially serve multiple customer classes and can allow for server idling in the presence of waiting customers. Each WRBS policy has associated with it a probability vector

$p := (p_1, \ldots, p_J)$, where J is the number of customer classes. Under WRBS policy p, the chance that a newly available server next serves class j is p_j, assuming there is a class j customer waiting. To handle cases where there are no customers of the chosen class waiting, we specify a protocol for determining when to choose an alternative class and when to idle within the definition of WRBS in Section 5.1. For time homogeneous arrival processes, the specifics of this protocol should not profoundly affect the overall system behavior, provided the probability vector p is chosen appropriately (i.e., is in line with the workload arriving from each class). In that case, the frequency with which each class gets called into service when there are no jobs of that class waiting will be asymptotically negligible and the WRBS policy will perform asymptotically equivalently to a nonidling policy in the fluid limit (see the discussion in Section 5.2.3). To support this, we specify a WRBS fluid model and classify the invariant states for the WRBS fluid model (see Theorem 4).

From a methodological standpoint, the analysis can be more or less complicated depending on what protocol is implemented when a server attempts to serve a class with no customer waiting. We specify this in a way that allows us to define a multidimensional regulator mapping having simple reflection directions (see Definition 7) that are Lipschitz continuous. In our analysis, this regulator mapping is the key to proving the uniqueness of WRBS fluid model solutions (see Theorem 3 and, more specifically, Lemma 13). Then, in light of Theorems 1, 2, and 3, in order to prove a fluid limit theorem, it suffices to show that a fluid limit point arising from a sequence of multiclass many-server queues operating under a WRBS policy satisfies the WRBS specific fluid relations. Under suitable asymptotic conditions, we prove this here as Theorem 5. The regulator mapping plays an integral role in the proof of Theorem 5 as well.

1.1.3. Organization of the Paper. The remainder of this paper is organized as follows. We end this section with a brief review of related literature and a subsection that summarizes our mathematical notation. In Section 2, the many-server queueing model is specified together with defining an HL control policy and an admissible HL control policy. Section 3 provides the fluid model equations relevant for any HL control policy and summarizes some properties of solutions to the fluid model equations. Section 4 contains our non-policy-specific results, a convergence result in Theorem 1, and an associated tightness result in Theorem 2. In Section 5, we take up the study of the WRBS policy class. There we formally define WRBS and provide policy-specific equations and inequalities that uniquely specific the system dynamics. We also provide the supplemental WRBS fluid model equations and inequalities, prove a uniqueness result for WRBS fluid model solutions (see Theorem 3), and classify the WRBS invariant states (see Theorem 4). Finally, we prove a WRBS fluid limit theorem (see Theorem 5).

1.1.4. Most Closely Related Literature. The mathematical machinery that we use for our fluid limit proofs builds on that developed in the single-class setting, which we review in this paragraph (see also the survey papers Dai and He [7] and Ward [20]). In Whitt [22], the author proposes a fluid approximation of an overloaded many-server $G/GI/N+GI$ queue; later work, Liu and Whitt [16], develops an algorithm to calculate performance functions in the more general case that both the arrival rate and service capacity are time varying. In Kang and Ramanan [11] and Zhang [25], the authors prove the convergence to a fluid model related to the one proposed in Whitt [22] under different assumptions on the service time distribution. The proofs in Kang and Ramanan [11] follow the approach in Kaspi and Ramanan [12] in the traditional many-server setting that does not allow for customers to abandon. In Atar et al. [4], this methodology is extended to the multiclass many server under static priority setting. Our generalization to the multiclass many-server queue with control setting identifies generic conditions sufficient to imply tightness (see Theorem 2). We also separate the conditions required to show that a distributional limit point of a sequence of scaled state processes satisfies the fluid model equations from the more restrictive conditions required to show weak convergence under a WRBS policy (see Theorems 1 and 5).

In the tutorial paper Puha and Ward [19, sections 4 and 5], we discuss the fluid model studied here. In Puha and Ward [19, sections 4 and 5] for time-stationary input, we classify the set of invariant states and propose a fluid control problem associated with the set of invariant states. Under suitable asymptotic conditions, such an optimization problem is expected to arise as the fluid limit of a certain long-run average cost function that penalizes for holding and reneging. In Puha and Ward [19, sections 5 and 6], we show that when only reneging is penalized, the solution to the proposed fluid control problem suggests that a static priority policy is asymptotically optimal. However, when reneging and holding are both penalized, the solution to the proposed fluid control problem suggests that a static priority policy is not necessarily asymptotically optimal (Puha and Ward [19, lemma 1 and remark 10]), which is in contrast to the asymptotic optimality results in Atar et al. [3] for the multiclass many-server M/M/N+M queue and in Atar et al. [4] for the multiclass many-server G/GI/N+M queue. In work in progress, we aim to prove an asymptotic optimality result under fully general distributional assumptions for the optimization problem introduced in Puha and Ward [19]; see also Long et al. [17] for work in this direction in a G/M/N+GI setting.

1.1.5. Notation. The following notation will be used throughout this paper. We denote the set of integers by $\mathbb{Z}$, the set of positive integers by $\mathbb{N}$, the set of nonnegative integers by $\mathbb{Z}_+$, the set of nonpositive integers by $\mathbb{Z}_-$, the set of real numbers by $\mathbb{R}$, and the set of nonnegative real numbers by $\mathbb{R}_+$. For $a, b \in \mathbb{R}$, both $a \vee b$ and $\max(a,b)$ (respectively $a \wedge b$ and $\min(a,b)$) denote the maximum (resp. minimum) of a and b. Also, the shorthand a^+ and a^- are used for $a \vee 0$ and $-a \vee 0$, respectively, and $|a| = a^+ \vee a^-$. The sets $\mathbb{R}$ and $\mathbb{R}_+$ are endowed with the Euclidean topology, and $\mathbb{Z}$ and $\mathbb{Z}_+$ are endowed with the discrete topology.

For sets A and B, $A \times B$ denotes the Cartesian product of A and B. If A and B are topological spaces, $A \times B$ is the product space endowed with the product topology. For $k \in \mathbb{N}$ and a set S, S^k denotes the k-fold Cartesian product $S \times S \times \cdots \times S$ and for a topological space $\mathbb{S}$, $\mathbb{S}^k$ is the k-fold Cartesian product endowed with the product topology. For $k \in \mathbb{N}$ and $x \in \mathbb{R}^k$, $\|x\| = \sum_{i=1}^k |x_i|$ and $x_\Sigma = \sum_{i=1}^k x_i$. Then, $\|\cdot\|$ denotes the l^1-norm on $\mathbb{R}^k$, where the particular $k \in \mathbb{N}$ will be clear from context.

For a measurable space $(S, \mathcal{F})$ and a measurable set $A \in \mathcal{F}$, 1_A is the indicator function of the set A, which is one when its argument is a member of the set A and is zero otherwise. In addition, when A is S, we use the shorthand notation 1 to mean 1_S. Also, for $\mathcal{A} \subset \mathcal{F}$, $\sigma(\mathcal{A})$ denotes the σ-algebra generated by $\mathcal{A}$.

Let $H \in (0, \infty]$. For a Borel measurable $\varphi : [0,H) \times \mathbb{R}_+ \to \mathbb{R}$, $\mathrm{supp}(\varphi) \subseteq [0,H) \times \mathbb{R}_+$ denotes the support of φ and $\|\varphi\|_\infty = \sup\{|\varphi(x,t)| : (x,t) \in [0,H) \times \mathbb{R}_+\}$. Also, $\mathbf{C}_c([0,H))$ (resp. $\mathbf{C}_b([0,H))$) denotes the set of continuous, compactly supported (resp. bounded) functions $f : [0,H) \to \mathbb{R}$; and $\mathbf{C}_c^1([0,H))$ denotes the set of continuous, compactly supported functions $f : [0,H) \to \mathbb{R}$ for which the derivative f' exists for all $x \in [0,H)$ and $t \geq 0$ and lies in $\mathbf{C}_c([0,H))$. Similarly, $\mathbf{C}_c([0,H) \times \mathbb{R}_+)$ (resp. $\mathbf{C}_b([0,H) \times \mathbb{R}_+)$) denotes the set of continuous, compactly supported (resp. bounded) functions $\varphi : [0,H) \times \mathbb{R}_+ \to \mathbb{R}$; $\mathbf{C}_c^{1,1}([0,H) \times \mathbb{R}_+)$ denotes the set of continuous, compactly supported functions $\varphi : [0,H) \times \mathbb{R}_+ \to \mathbb{R}$ for which the directional derivative $\lim_{\varepsilon \to 0} \frac{\varphi(x+\varepsilon, t+\varepsilon) - \varphi(x,t)}{\varepsilon}$ exists for all $x \in [0,H)$ and $t \geq 0$ and lies in $\mathbf{C}_c([0,H), \mathbb{R}_+)$. We shall abuse the notation by using $\varphi_x + \varphi_t$ to denote this directional derivative, whether the partial derivatives exist or not. Finally, $\mathbf{L}^1([0,H))$ (resp. $\mathbf{L}^1_{\mathrm{loc}}([0,H))$) denotes the set of Borel measurable functions on $[0,H)$ that are integrable (resp. locally integrable) with respect to Lebesgue measure on $[0,H)$.

Given a Polish space $\mathbb{S}$, we use the notation $\mathbf{D}(\mathbb{S})$ to denote the set of $\mathbb{S}$ valued functions of $\mathbb{R}_+$ that are right continuous with finite lefts (rcll), endowed with the usual Skorokhod J_1-topology (Billingsley [6]). In contrast to the sets of functions defined in the previous paragraph denoted using $\mathbf{C}$, we use the range rather than the domain as the argument because the domain is always time, $\mathbb{R}_+$. For $f \in \mathbf{D}(\mathbb{S})$, $f(0-) = f(0)$ and $f(t-) = \lim_{u \nearrow t} f(u)$ for all $t > 0$. In addition, when $\mathbb{S} = \mathbb{R}_+^k$ for some $k \in \mathbb{N}$, we define $\mathbf{D}_\uparrow(\mathbb{R}_+^k)$ to be the set of members of $\mathbf{D}(\mathbb{R}_+^k)$ such that each coordinate is nondecreasing and has initial value zero. For $k \in \mathbb{N}$, $f \in \mathbf{D}(\mathbb{R}_+^k)$ and $t \geq 0$, $\|f\|_t = \sup_{0 \leq u \leq t} \|f(u)\|$. Finally, 0 denotes the process in $\mathbf{D}(\mathbb{R})$ that is identically equal to zero. All processes considered in this paper are assumed to be rcll, unless explicitly otherwise indicated.

Let $H \in (0, \infty]$. Then $\mathbf{M}[0,H)$ denotes the set of finite, nonnegative Borel measures on $[0,H)$ endowed with the topology of weak convergence, which is a Polish space. For given $\eta \in \mathbf{M}[0,H)$ and a Borel measurable function $f : [0,H) \to \mathbb{R}_+$ that is integrable with respect to η, we let $\langle f, \eta \rangle = \int_{[0,H)} f(x) \eta(dx)$. Given $x \in [0,H)$, $\delta_x \in \mathbf{M}[0,H)$ denotes the Dirac measure with unit atom at x, that is, for all Borel measurable $A \subset [0,H)$, $\langle 1_A, \delta_x \rangle = 1_A(x)$. Then $\mathbf{M}_D[0,H)$ denotes the subset of $\mathbf{M}[0,H)$ consisting of the measures in $\mathbf{M}[0,H)$ that can be represented as a sum of finitely many Dirac measures, whereas $\mathbf{M}_{D_1}[0,H)$ denotes the subset of $\mathbf{M}_D[0,H)$ that consists of those measures in $\mathbf{M}_D[0,H)$ that can be represented as a sum of finitely many distinct Dirac measures. As shown in Lee [15, appendix B.1], $\mathbf{M}_{D_1}[0,H)$ endowed with the topology of weak convergence is a Polish space. By adapting the argument in Lee [15, appendix B.1] to account for a finite number of Dirac measures coinciding, it can be shown that $\mathbf{M}_D[0,H)$ endowed with the topology of weak convergence is also a Polish space. Finally, we say that $\eta \in \mathbf{M}[0,H)$ does not charge points if $\langle 1_{\{x\}}, \eta \rangle = 0$ for all $x \in [0,H)$.

Given a cumulative distribution function G defined on $[0, \infty]$ that is absolutely continuous on $\mathbb{R}_+$ with density function g, we define some quantities associated with G. The right edge of the support of G is given by

$$H = \sup\{x \in \mathbb{R}_+ : G(x) < 1\}. \tag{1}$$

Then $H \in (0, \infty]$. The hazard function h is given by

$$h(x) = \frac{g(x)}{1 - G(x)}, \qquad x \in [0,H).$$

Note that the hazard function $h \in \mathbf{L}^1_{\mathrm{loc}}([0,H))$. To see this, note that by assumption G is absolutely continuous on $\mathbb{R}_+$ and, because $\ln(\cdot)$ is Lipschitz continuous on $[a, \infty)$ for any $a > 0$, it follows that $-\ln(1 - G(x))$, $x \in [0,H)$ is

absolutely continuous on $[0,b]$ for any $b < H$. Furthermore, h is the almost everywhere derivative of $-\ln(1-G(\cdot))$ on $[0,H)$. Thus, for all $0 \le x < y < H$,

$$\int_x^y h(t)dt = -\ln(1-G(y)) - (-\ln(1-G(x))) < \infty,$$

(see Folland [9, corollary 3.34 and theorem 3.36]). Given $y \in [0,H)$, let G_y be the conditional cumulative distribution function conditioned to exceed y. In particular, given $y \in [0,H)$,

$$G_y(x) = \frac{G(x+y) - G(y)}{1 - G(y)}, \qquad x \in [0,\infty]. \tag{2}$$

2. Multiclass *N*-Server Queues with Control

We consider a sequence of multiclass many-server queues, indexed by elements of $\mathbb{N}$. The number of customer classes $J \in \mathbb{N}$ is fixed throughout. We set $\mathbb{J} = \{1,2,\dots,J\}$. The queue indexed by $N \in \mathbb{N}$ has N identical servers and J customer classes and is referred to as the N-server multiclass queue or the N-server queue for shorthand. For a given $N \in \mathbb{N}$, customers enter each class exogenously requiring a random amount of service that can be processed by any one of the N identical servers. Each customer also has associated with it a patience time of random length, which could be infinite and indicates how long that customer is willing to wait in system to begin service, prior to abandoning the queue. More specifically, a customer's potential abandonment time is his or her arrival time plus his or her patience time. Customers that do not enter service between their arrival time and potential abandonment time abandon the system at their potential abandonment time. Customers can arrive individually or in batches. New customers that arrive to find all servers busy must wait in the queue for their class; otherwise, the control policy decides how many customers enter service. The control policies considered here are head-of-the-line within each class. When any given server becomes available, the control policy determines when that server will begin serving another customer and which customer class will be served. In particular, the control policy may elect to allow that server to idle for a time before sending a customer into service. At the moment when a class j customer is sent into service, the customer waiting the longest from that class is the one that enters service. At moments when there is more than one idle server, the control policy might elect to send multiple customers of the same or different classes into service simultaneously. The control policies considered here may be a deterministic function of the system state or may invoke some randomness.

In what follows, we formally define the stochastic model. Section 2.1 provides the primitive model inputs, and Section 2.2 sets up the state space. We separate the system dynamics into those that are independent of the control policy, shown in Section 2.3, and those that depend on the control policy, shown in Section 2.4. In Section 2.5, we define the terminology HL control policy. The class of admissible HL control policies, defined in Section 2.6, is the class for which fluid limit points arising under certain asymptotic assumptions are almost surely fluid model solutions (see Theorem 1).

The sequence of N-server queues is defined on a fixed probability space $(\Omega,\mathcal{F},\mathbb{P})$. For the remainder of this section, $N \in \mathbb{N}$ is regarded as fixed. In preparation for taking limits as $N \to \infty$ in Section 4, we superscript all quantities that depend on N by N.

2.1. Primitive Inputs

Here, we define the arrival process, which dictates when customers enter the system, or arrive, and fix several sequences of independent and identically distributed (i.i.d.) random variables defined on $(\Omega,\mathcal{F},\mathbb{P})$ that serve to provide primitive inputs for the model.

2.1.1. The Arrival Process. For each $j \in \mathbb{J}$, let E_j^N denote a counting process. In particular, for each $j \in \mathbb{J}$, E_j^N is a nondecreasing, pure jump process with jump sizes taking values in $\mathbb{N}$ such that $E_j^N(0) = 0$ and $E_j^N(t) < \infty$ for all $t \ge 0$, almost surely. Then, for each $j \in \mathbb{J}$ and $t \ge 0$, $E_j^N(t)$ denotes the number of class j customers to enter the system, or arrive, in $(0,t]$. Because the jump sizes may be larger than one, a finite number of customers may arrive simultaneously. We refer to a collection of customers that arrive simultaneously as a batch. For each $j \in \mathbb{J}$ and $i \in \mathbb{N}$, let

$$e_{j,i}^N = \inf\{t \ge 0 : E_j^N(t) \ge i\},$$

which is the time at which the i^{th} class j customer arrives; we refer to such a customer as the i^{th} class j arrival. We assume that $\mathbb{P}\left(e_{j,1}^N < \infty\right) = 1$.

For each $j \in \mathbb{J}$, let $\alpha^N_{0,j}$ be a random variable taking values in $\mathbb{R}_+$. For each $j \in \mathbb{J}$, $\alpha^N_{0,j}$ is interpreted as the time that has elapsed by time zero since the most recent class j batch arrived among those batches that entered the system at or prior to time zero. For each $j \in \mathbb{J}$ and $t \geq 0$, set

$$\alpha^N_j(t) = \begin{cases} \alpha^N_{0,j} + t, & 0 \leq t < e^N_{j,1}, \\ t - \sup\{s < t : E^N_j(t) - E^N_j(s) > 0\}, & t \geq e^N_{j,1}. \end{cases} \tag{3}$$

Then, for each $j \in \mathbb{J}$ and $t \geq 0$, $\alpha^N_j(t)$ denotes the time that has elapsed by time t since the most recent class j batch arrived among those batches that entered the system at or prior to time t. It is assumed that α^N_j is Markovian with respect its own natural filtration for each $j \in \mathbb{J}$. This holds if E^N_j is a renewal process, in which case α^N_j is the backward recurrence time process associated with E^N_j, or if E^N_j is a time inhomogeneous Poisson process.

Let E^N (resp. α^N) denote the vector process with j^{th} coordinate E^N_j (resp. α^N_j) for $j \in \mathbb{J}$. It is assumed that the coordinates of E^N are mutually independent.

2.1.2. Additional Primitive Inputs. For each $j \in \mathbb{J}$, $\{v_{j,i}\}_{i\in\mathbb{Z}}$ is an i.i.d. sequence of positive random variables, used to represent service times, defined on $(\Omega, \mathcal{F}, \mathbb{P})$ that have common absolutely continuous cumulative distribution function G^s_j on $\mathbb{R}_+$ with probability density function g^s_j. We let H^s_j be the right edge of the support of G^s_j, h^s_j be the associated hazard function, and $G^s_{j,y}$ the conditional cumulative distribution function conditioned to exceed y for $y \in [0, H^s_j)$.

For each $j \in \mathbb{J}$, $\{r_{j,i}\}_{i\in\mathbb{N}}$ is an i.i.d. sequence of positive random variables, used to represent patience times, defined on $(\Omega, \mathcal{F}, \mathbb{P})$ that have common cumulative distribution function G^r_j on $[0, \infty]$ that, when restricted to $\mathbb{R}_+$, has probability density function g^r_j. In contrast to service times, potential abandonment times can be infinite; our model reduces to a standard multiclass $G/GI/N$ queue when all abandonment times are infinite. We let H^r_j be the right edge of the support of G^r_j, h^r_j be the associated hazard function, and $G^r_{j,y}$ the conditional cumulative distribution function conditioned to exceed y for $y \in [0, H^r_j)$.

For each $j \in \mathbb{J}$, let $\{V^N_{j,i}\}_{i\in\mathbb{Z}_-}$ and $\{R^N_{j,i}\}_{i\in\mathbb{Z}_-}$ be collections of i.i.d. uniform (0, 1) random variables. These will be used below to define various random residual times associated with the initial condition. In addition, let $\{\epsilon^N_i\}_{i\in\mathbb{N}}$ and $\{d^N_i\}_{i\in\mathbb{N}}$ be i.i.d. sequences of uniform (0, 1) random variables. These may be used if the control policy invokes some randomness, as we will see.

The sequences $\{V^N_{j,i}\}_{i\in\mathbb{Z}_-}$, $j \in \mathbb{J}$, $\{R^N_{j,i}\}_{i\in\mathbb{Z}_-}$, $j \in \mathbb{J}$, $\{v_{j,i}\}_{i\in\mathbb{Z}}$, $j \in \mathbb{J}$, $\{r_{j,i}\}_{i\in\mathbb{N}}$, $j \in \mathbb{J}$, $\{\epsilon^N_i\}_{i\in\mathbb{N}}$, and $\{d^N_i\}_{i\in\mathbb{N}}$ are all assumed to be mutually independent of one another and of E^N. We refer to this collection of random sequences together with E^N as the stochastic primitive inputs, or simply the primitive inputs, for the N-server queue.

2.2. The State Space

The N-server queue with control is an rcll process taking values in the set

$$\mathbb{Y} = \mathbb{R}^J_+ \times \mathbb{Z}^J_+ \times \times^J_{j=1} \mathbf{M}_D[0, H^s_j) \times \times^J_{j=1} \mathbf{M}_D[0, H^r_j). \tag{4}$$

Given $y \in \mathbb{Y}$, we write $y = (\alpha, x, \nu, \eta)$, where $\alpha \in \mathbb{R}^J_+$, $x \in \mathbb{Z}^J_+$, $\nu \in \times^J_{j=1} \mathbf{M}_D[0, H^s_j)$, and $\eta \in \times^J_{j=1} \mathbf{M}_D[0, H^r_j)$. Because $\mathbb{Y}$ is a product of Polish spaces, $\mathbb{Y}$ is a Polish space.

We begin by informally explaining what each coordinate of $y \in \mathbb{Y}$ represents in terms of the N-server queue. Given $y = (\alpha, x, \nu, \eta)$, for each $j \in \mathbb{J}$, $\alpha_j \in \mathbb{R}_+$ is the time that has elapsed since the last class j batch of customers arrived to the system (as in Section 2.1.1) and $x_j \in \mathbb{Z}_+$ is the number of class j customers in system. For each $j \in \mathbb{J}$, ν_j is a measure in $\mathbf{M}_D[0, H^s_j)$ that has a unit mass at the age-in-service (amount of service received) of each class j customer currently in service. In particular, $\langle 1, \nu_j \rangle$ is the total mass of ν_j, which denotes the number of class j customers currently in service. Then $q_j = x_j - \langle 1, \nu_j \rangle$ denotes the number of class j customers currently in system waiting for service. We refer to such customers as customers in queue. For each $j \in \mathbb{J}$, η_j is a measure in $\mathbf{M}_D[0, H^r_j)$ that has a unit mass at the potential waiting time of each customer "potentially" in system. Customers "potentially" in system are those that have entered the system, 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 what follows, we define these objects precisely. But first we must restrict the state space in order to respect certain natural constraints of the N-server queue.

The system state will be an element of $\mathbb{Y}$ for all time and will satisfy some additional constraints, some of which depend on N. For this, let $\mathbb{Y}^N$ be the subset of $y^N = (\alpha^N, x^n, \nu^N, \eta^N) \in \mathbb{Y}$ such that

(S.1) for each $j \in \mathbb{J}$, $\langle 1, \nu_j^N \rangle \le x_j^N \le \langle 1, \nu_j^N \rangle + \langle 1, \eta_j^N \rangle$ and

(S.2) $\sum_{j=1}^{J} \langle 1, \nu_j^N \rangle \le N$.

A consequence of (S.1) is that the number of class j customers in queue is nonnegative and cannot exceed the number of potential class j customers, for each $j \in \mathbb{J}$. A consequence of (S.2) is that the total number of customers in service cannot exceed the number of servers. Because $\mathbb{Y}^N$ is a closed subset of (4), $\mathbb{Y}^N$ is a Polish space.

Throughout Sections 2.3 and 2.4, we fix $y_0^N \in \mathbb{Y}^N$, where $y_0^N = (\alpha_0^N, x_0^N, \nu_0^N, \eta_0^N)$ with the j^{th} coordinates being respectively denoted by $\alpha_{0,j}^N$, $x_{0,j}^N$, $\nu_{0,j}^N$, and $\eta_{0,j}^N$, for $j \in \mathbb{J}$. In addition, for convenience, for each $j \in \mathbb{J}$, we define

$$b_{0,j}^N = \langle 1, \nu_{0,j}^N \rangle \qquad \text{and} \qquad q_{0,j}^N = x_{0,j}^N - b_{0,j}^N.$$

Then, for each $j \in \mathbb{J}$, $b_{0,j}^N$, $q_{0,j}^N$, $x_{0,j}^N$, and $\langle 1, \eta_{0,j}^N \rangle$ respectively denote the number of class j customers in service, in queue, in system, and in the potential queue at time zero.

2.3. Control Policy Independent System Dynamics

In this section, we define the system dynamics and inputs that are derived from the initial state $y_0^N \in \mathbb{Y}^N$ and stochastic primitive inputs, that is, the system dynamics and inputs that do not depend on the control policy.

2.3.1. Patience Times and the Potential Queue Process. Fix a customer class $j \in \mathbb{J}$. There are $\langle 1, \eta_{0,j}^N \rangle$ class j potential customers that arrived at or prior to time zero whose potential abandonment time is after time zero. These are referred to as *time zero potential customers*, because some may have entered service and even departed the system by completing service prior to time zero. For each class $j \in \mathbb{J}$, we index the time zero potential customers by nonpositive integers in the order of their arrival. For this, if $\langle 1, \eta_{0,j}^N \rangle \ge 1$, let $\{w_{j,i}^N(0)\}_{i=-\langle 1, \eta_{0,j}^N \rangle + 1}^{0}$ denote the nonincreasing sequence of the locations of the $\langle 1, \eta_{0,j}^N \rangle$ Dirac measures that comprise $\eta_{0,j}^N$. In particular,

$$\eta_{0,j}^N = \sum_{i=-\langle 1, \eta_{0,j}^N \rangle + 1}^{0} \delta_{w_{j,i}^N(0)}, \tag{5}$$

where

$$0 \le w_{j,0}^N(0) \le w_{j,-1}^N(0) \le \cdots \le w_{j,-\langle 1, \eta_{0,j}^N \rangle + 1}^N(0) < H_j^r. \tag{6}$$

For $i = -\langle 1, \eta_{0,j}^N \rangle + 1, \ldots, 0$, $w_{j,i}^N(0)$ corresponds to the amount of time that has elapsed since class j time zero potential customer i arrived. Because customers may arrive in a batch, the equalities in (6) are weak. For $i = -\langle 1, \eta_{0,j}^N \rangle + 1, \ldots, 0$, we set

$$e_{j,i}^N = -w_{j,i}^N(0),$$

which is nonpositive and denotes the actual time at which class j time zero potential customer i arrived. Recall that $q_{0,j}^N \le x_{0,j}^N \wedge \langle 1, \eta_{0,j}^N \rangle$ and there are $q_{0,j}^N$ class j customers in queue at time zero. For $i = -q_{0,j}^N + 1, \ldots, 0$, we also refer to potential class j customer i as the i^{th} class j arrival.

For each $i \in \mathbb{N}$, $r_{j,i}$ denotes the patience time of the i^{th} class j arrival. In particular, if the i^{th} class j arrival does not enter service in the time interval $[e_{j,i}^N, e_{j,i}^N + r_{j,i})$, then that customer abandons the queue at time $e_{j,i}^N + r_{j,i}$. For $i \in \mathbb{N}$, set $r_{j,i}^N = r_{j,i}$. For $i = -\langle 1, \eta_{0,j}^N \rangle + 1, \ldots, 0$, class j time zero potential customer i entered the system at time $e_{j,i}^N$ ($w_{j,i}^N(0)$ time units prior to time zero) and did not request to abandon the system by time zero. Therefore, the patience time of this customer necessarily exceeds $w_{j,i}^N(0)$. For $i = -\langle 1, \eta_{0,j}^N \rangle + 1, \ldots, 0$, the patience time $r_{j,i}^N$ of class j time zero potential customer i is given by

$$r_{j,i}^N = \inf\{t > 0 : G_{j, w_{j,i}^N(0)}^r(t) > R_{j,i}^N\} + w_{j,i}^N(0).$$

For each $i = -\langle 1, \eta_{0,j}^N \rangle + 1, \ldots, 0, 1, 2, \ldots$ and $t \ge 0$, the potential waiting time of class j potential customer i at time t is given by

$$w_{j,i}^N(t) = \min\left(\left[t - e_{j,i}^N\right]^+, r_{j,i}^N\right).$$

The class j potential queue process η_j^N takes values in $\mathbf{M}_D[0, H_j^r)$. For $t \geq 0$,

$$\eta_j^N(t) = \sum_{j=-\langle 1, \eta_{0,j}^N \rangle + 1}^{E_j^N(t)} \delta_{w_{j,i}^N(t)} 1_{\{0 \leq t - e_{j,i}^N < r_{j,i}^N\}}. \tag{7}$$

Then, for each $t \geq 0$, $\langle 1, \eta_j^N(t) \rangle$ is the number of class j potential customers in the queue that arrived by time t and whose potential waiting time is less than their patience time. Note that at time t such customers may be in queue, in service, or may have been served to completion and departed the system.

2.3.2. Service Times. Fix a customer class $j \in \mathbb{J}$. Recall that there are $b_{0,j}^N = \langle 1, \nu_{0,j}^N \rangle$ class j customers in service at time zero and $q_{0,j}^N = x_{0,j}^N - b_{0,j}^N$ customers in queue at time zero. We index class j customers in service at time zero using the nonpositive integers $-x_{0,j}^N + 1, \ldots, -q_{0,j}^N$ in the order of time spent in service, with the class j customer that has been in service for the longest time being associated with the smallest index $-x_{0,j}^N + 1$. We index class j customers in queue at time zero using the nonpositive integers $-q_{0,j}^N + 1, \ldots, 0$ in the order of time spent waiting in queue, with the class j customer that has been waiting in the system for the longest time being associated with the smallest index $-q_{0,j}^N + 1$.

Recall that the indices $i = -q_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots$ correspond to class j arrivals that have not entered service by time zero. For $i = -q_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots$, the random variable $v_{j,i}$ denotes the service time requirement of the i^{th} class j arrival. In particular, if the i^{th} class j arrival enters service at some point in time, it will remain in service for exactly $v_{j,i}$ units of time. For $i = -q_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots$, set $v_{j,i}^N = v_{j,i}$.

The measure $\nu_{0,j}^N$ encodes the age-in-service of each class j customer in service at time zero. Specifically, let $\{a_{j,i}^N(0)\}_{i=-x_{0,j}^N+1}^{-q_{0,j}^N}$ denote the nondecreasing sequence of the locations of the unit Dirac measures of $\nu_{0,j}^N$. Then

$$\nu_{0,j}^N = \sum_{i=-x_{0,j}^N+1}^{-q_{0,j}^N} \delta_{a_{j,i}^N(0)},$$

where

$$0 \leq a_{j,-q_{0,j}^N}^N(0) \leq a_{j,-q_{0,j}^N-1}^N(0) \leq \cdots \leq a_{j,-x_{0,j}^N+1}^N(0) < H_j^s. \tag{8}$$

For $i = -x_{0,j}^N + 1, \ldots, -q_{0,j}^N$, the i^{th} class j arrival is in service at time zero and has received $a_{j,i}^N(0)$ units of service by time zero. We will consider control policies that allow more than one customer to enter service simultaneously, and therefore the inequalities in (8) are weak. The service times for these customers are defined as follows: for $i = -x_{0,j}^N + 1, \ldots, -q_{0,j}^N$, set

$$v_{j,i}^N = \inf\{t > 0 : G_{j,a_{j,i}^N(0)}^s(t) > V_{j,i}^N\} + a_{j,i}^N(0).$$

Note that, for $i = -q_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots$, class j potential customer i and arrival i are the same entity, and often we will simply refer to this entity as class j customer i or the i^{th} class j customer. However, for $i = -(x_{0,j}^N \vee \langle 1, \eta_{0,j}^N \rangle) + 1, \ldots, -q_{0,j}^N$, this is not necessarily the case. This is because class j customers in service at time zero may have virtually reneged after entering service, but prior to time zero. In this case, such customers are not associated with a Dirac measure in $\eta_{0,j}^N$. Similarly, class j time zero potential customers may have entered and completed service by time zero; such potential customers are not associated with a Dirac measure in $\nu_{0,j}^N$.

Also, note that the customer arrival times and quantities associated with time zero customers or potential customers are superscripted with N and may change with N. The service times of arrivals that enter service after time zero and the patience times of potential customers that arrive after time zero do not change with N. However, we superscript them with N in order to simplify the representation of future expressions.

2.4. Control Policy Dependent System Dynamics

The control policies considered here are HL within each class (and are defined precisely in Definition 4). For this given $j \in \mathbb{J}$, we make the convention that for $i \in \mathbb{N}$, the i^{th} HL class j customer is the class j customer in queue that, among all class j customers in queue, has the i^{th} smallest index, if such a customer exists, that is, if the number of class j customers in queue is at least i. Because indices are assigned to customers in the order of their arrival, the

i^{th} HL class j customer has the i^{th} largest potential waiting time among all class j customers in queue. The phrase the HL class j customer is often used as a shorthand for the 1^{st} HL class j customer. In addition to being HL within each class, the control policies considered here allow more than one customer to enter service simultaneously. Then, when $k \in \mathbb{N}$ class j customers simultaneously enter service, they are the 1^{st} through k^{th} HL class j customers in queue, that is, the k class j customers that have been waiting the longest. This is made precise in (10). It is further assumed that service is nonpreemptive and that customers are served one at a time by a single server dedicated to processing the work associated with that customer. In particular, there is no resource sharing or simultaneous resource possession. This is encapsulated in (11). In this section, we develop the natural conditions that should be satisfied by a many-server queue operating under such an HL control policy.

2.4.1. The Entry-Into-Service Process. The control policy must determine how many customers from each class enter service by each time $t > 0$. In particular, a state descriptor for the N-server queue with initial value $y_0^N \in \mathbb{Y}^N$ is a $\mathbb{Y}^N$ valued process Y^N for which $Y^N(0) = y_0^N$; there is an associated $\mathbb{Z}_+^J$ valued process K^N such that, for each $j \in \mathbb{J}$, K_j^N is a counting process with initial value zero, where $K_j^N(t)$ denotes the number of class j customers that enter service in $(0,t]$, for $t \geq 0$. Here we record some natural conditions that such a process K^N should satisfy.

Fix a customer class $j \in \mathbb{J}$. For $i = -x_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots, k_{j,i}^N$ denotes the time of entry into service of the i^{th} class j customer. It is set to infinity if that customer abandons the queue before entering service. For $i = -x_{0,j}^N + 1, \ldots, -q_{0,j}^N$, it follows that $k_{j,i}^N = -a_{j,i}^N(0)$. For $i = -q_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots, k_{j,i}^N$ is determined by the system dynamics and the HL control policy. For $i = -q_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots$ such that $k_{j,i}^N < \infty$, it is assumed that

$$e_{j,i}^N \leq k_{j,i}^N < e_{j,i}^N + r_{j,i}^N, \tag{9}$$

which enforces that customers cannot enter service prior to arriving to the system or once they abandon the system. It is also assumed that, for all $-x_{0,j}^N + 1 \leq i < l < \infty$ such that $k_{j,i}^N \vee k_{j,l}^N < \infty$,

$$k_{j,i}^N \leq k_{j,l}^N. \tag{10}$$

This implies that the class j customers that do not abandon the queue prior to entering service are served in the order of their time of arrival, that is, in HL fashion. It follows that, for each $t \geq 0$, the number $K_j^N(t)$ of class j customers to enter service in the time interval $(0,t]$ is given by

$$K_j^N(t) = \sum_{i=-q_{0,j}^N+1}^{E_j^N(t)} 1_{\{k_{j,i}^N \leq t\}}.$$

Then $K_j^N(0) = 0$. We refer to the J-dimensional vector process K^N with i^{th} coordinate K_i^N, $1 \leq i \leq J$, as the entry-into-service process.

2.4.2. The Age-in-Service Process. Fix a customer class $j \in \mathbb{J}$. For each $i = -x_{0,j}^N + 1, \ldots, 0, 1, 2, \ldots$, and $t \geq 0$, the age-in-service $a_{j,i}^N(t)$ of the i^{th} class j customer at time $t \geq 0$ is given by

$$a_{j,i}^N(t) = \min\left(\left[t - k_{j,i}^N\right]^+, v_{j,i}^N\right), \tag{11}$$

which corresponds to the amount of service received by time t. In particular, once a customer enters service, that customer is served continuously at rate one until it receives its full service requirement, at which time the customer exits the system. The class j age-in-service process ν_j^N takes values in $\mathbf{M}_D[0, H_j^s)$. For $t \geq 0$,

$$\nu_j^N(t) = \sum_{i=-x_{0,j}^N+1}^{E_j^N(t)} \delta_{a_{j,i}^N(t)} 1_{\{0 \leq t - k_{j,i}^N < v_{j,i}^N\}}. \tag{12}$$

Then, for $t \geq 0$, the number of class j customers that are in service at time t is given by

$$B_j^N(t) = \langle 1, \nu_j^N(t) \rangle. \tag{13}$$

We refer to B as the busy server process. For convenience later on, for $t \geq 0$, we let $\tilde{\nu}_j^N(t) \in \mathbf{M}_D[0, H_j^s)$ be given by

$$\tilde{\nu}_j^N(t) = \sum_{i=-x_{0,j}^N+1}^{E_j^N(t)} \delta_{a_{j,i}^N(t)} 1_{\{0 < t - k_{j,i}^N < v_{j,i}^N\}}, \qquad t \geq 0. \tag{14}$$

In particular, if a class j departure occurs at time t, then the unit atom that had been associated with that customer immediately prior to time t is not included in $\tilde{\nu}_j^N(t)$, in the same way that it is not included in $\nu_j^N(t)$. However, if a class j customer enters service at time t, then $\tilde{\nu}_j^N(t)$ neglects to include a unit atom at the origin corresponding to that customer, whereas $\nu_j^N(t)$ includes this unit atom. In other words, $\tilde{\nu}_j^N$ records departures at the moment at which they happen and entry-into-service the moment immediately after it happens. So then $\tilde{\nu}_j^N(t)$ differs from $\nu_j^N(t)$ only at times t that are jump times of K_j^N, and $\tilde{\nu}_j^N$ is not an rcll process.

2.4.3. Auxiliary Processes and System Evolution Equations. For each customer class $j \in \mathbb{J}$ and $t \geq 0$, $X_j^N(t)$ denotes the number of class j customers in system at time t and $Q_j^N(t)$ denotes the number of class j customers in queue at time t. As suggested by mass balance, we require that, for each $j \in \mathbb{J}$ and $t \geq 0$,

$$Q_j^N(t) = X_j^N(t) - B_j^N(t) \geq 0, \tag{15}$$

which is consistent with (S.1) because of (13). For $t \geq 0$, $I^N(t)$ denotes the number of idle servers, which is given by

$$I^N(t) = N - \sum_{j=1}^{J} B_j^N(t) \geq 0, \tag{16}$$

which is consistent with (S.2). We refer to I^N as the idle server process. For each customer class $j \in \mathbb{J}$, the class j departure process D_j^N, potential reneging process S_j^N, and reneging process R_j^N are given as follows: for each $t \geq 0$,

$$D_j^N(t) = \sum_{i=-x_{0,j}^N+1}^{E_j^N(t)} \sum_{s\in[0,t]} 1_{\left\{\frac{da_{j,i}^N}{dt}(s-)>0, \frac{da_{j,i}^N}{dt}(s+)=0\right\}},$$

$$S_j^N(t) = \sum_{i=-\langle 1,\eta_j^N(0)\rangle+1}^{E_j^N(t)} \sum_{s\in[0,t]} 1_{\left\{\frac{dw_{j,i}^N}{dt}(s-)>0, \frac{dw_{j,i}^N}{dt}(s+)=0\right\}},$$

$$R_j^N(t) = \sum_{i=-q_{0,j}^N+1}^{E_j^N(t)} \sum_{s\in[0,t]} 1_{\left\{a_{j,i}^N(s)=0, \frac{dw_{j,i}^N}{dt}(s-)>0, \frac{dw_{j,i}^N}{dt}(s+)=0\right\}}.$$

In particular, class j arrivals that abandon when their age-in-service is zero are regarded as not having entered service before the moment that they abandon. Hence, such customers are counted by the reneging process. Because of (9), such customers never enter service.

The system balance equations are given as follows: for each $j \in \mathbb{J}$ and $t \geq 0$,

$$X_j^N(t) = X_j^N(0) + E_j^N(t) - D_j^N(t) - R_j^N(t), \tag{17}$$

$$\langle 1, \eta_j^N(t)\rangle = \langle 1, \eta_j^N(0)\rangle + E_j^N(t) - S_j^N(t), \tag{18}$$

$$K_j^N(t) = B_j^N(t) + D_j^N(t) - B_j^N(0). \tag{19}$$

Upon subtracting $B_j^N(t)$ from both sides of (17) and subtracting and adding $B_j^N(0)$ on the right side of (17), using (15) twice and using (19), we find that for all $j \in \mathbb{J}$ and $t \geq 0$,

$$Q_j^N(t) = Q_j^N(0) + E_j^N(t) - R_j^N(t) - K_j^N(t). \tag{20}$$

Thus, using nonnegativity of Q_j^N, we find an implicit constraint on the entry-into-service-process, which says that $K_j^N(t) \leq Q_j^N(0) + E_j^N(t) - R_j^N(t)$ for all $t \geq 0$ and $j \in \mathbb{J}$.

Collections of marked point processes are used to characterize the dynamic evolution of the age-in-service process ν^N and potential reneging process η^N as follows. For each $j \in \mathbb{J}$, measurable function $\varphi : [0, H_j^s) \times \mathbb{R}_+ \to \mathbb{R}$ and measurable function $\psi : [0, H_j^r) \times \mathbb{R}_+ \to \mathbb{R}$, $\mathcal{D}_j^N(\varphi, \cdot)$ and $\mathcal{S}_j^N(\psi, \cdot)$ are the departure-from-service and potential

reneging marked point processes, defined as follows: for each $t \geq 0$,

$$\mathcal{D}_j^N(\varphi,t) = \sum_{i=-x_{0,j}^N+1}^{E_j^N(t)} \sum_{s\in[0,t]} 1_{\left\{\frac{da_{j,i}^N}{dt}(s-)>0, \frac{da_{j,i}^N}{dt}(s+)=0\right\}} \varphi(a_{j,i}^N(s),s),$$

$$\mathcal{S}_j^N(\psi,t) = \sum_{i=-\langle 1,\eta_j^N(0)\rangle+1}^{E_j^N(t)} \sum_{s\in[0,t]} 1_{\left\{\frac{dw_{j,i}^N}{dt}(s-)>0, \frac{dw_{j,i}^N}{dt}(s+)=0\right\}} \psi(w_{j,i}^N(s),s).$$

For each $j \in \mathbb{J}$, bounded measurable function $\varphi : [0, H_j^s) \times \mathbb{R}_+ \to \mathbb{R}$, bounded measurable function $\psi : [0, H_j^r) \times \mathbb{R}_+ \to \mathbb{R}$, and $t \geq 0$, we have

$$|\mathcal{D}_j^N(\varphi,t)| \leq \|\varphi\|_\infty \left(X_j^N(0) + E_j^N(t)\right), \tag{21}$$

$$|\mathcal{S}_j^N(\psi,t)| \leq \|\psi\|_\infty \left(\langle 1, \eta_j^N(0)\rangle + E_j^N(t)\right). \tag{22}$$

For $j \in \mathbb{J}$, a measurable function $f : [0, H_j^s) \to \mathbb{R}$ and a measurable function $\varsigma : [0, H_j^r) \to \mathbb{R}$, we note that if $\varphi_f(x,t) = f(x)$ for all $x \in [0, H_j^s)$ and $t \geq 0$ and $\psi_\varsigma(x,t) = \varsigma(x)$ for all $x \in [0, H_j^r)$ and $t \geq 0$, then φ_f and ψ_ς are measurable. Hence, for $j \in \mathbb{J}$, we adopt the shorthand notation $\mathcal{D}_j^N(f,\cdot) = \mathcal{D}_j^N(\varphi_f,\cdot)$ and $\mathcal{S}_j^N(\varsigma,\cdot) = \mathcal{S}_j^N(\psi_\varsigma,\cdot)$. Similarly, for $c \in \mathbb{R}_+$, $\mathcal{D}_j^N(c,\cdot) = cD_j^N(\cdot)$ and $\mathcal{S}_j^N(c,\cdot) = cS_j^N(\cdot)$.

Lemma 1. *For each $j \in \mathbb{J}$, $\varphi \in \mathbf{C}_c^{1,1}([0, H_j^s) \times \mathbb{R}_+)$, $f \in \mathbf{C}_c^1([0, H_j^s))$, $\psi \in \mathbf{C}_c^{1,1}([0, H_j^r) \times \mathbb{R}_+)$, $\varsigma \in \mathbf{C}_c^1([0, H_j^r))$, and $t \geq 0$,*

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_j^N(t)\rangle &= \langle \varphi(\cdot,0), \nu_j^N(0)\rangle + \int_0^t \langle \varphi_x(\cdot,u) + \varphi_t(\cdot,u), \nu_j^N(u)\rangle du \\ &\quad - \mathcal{D}_j^N(\varphi,t) + \int_0^t \varphi(0,u) dK_j^N(u),\end{aligned} \tag{23}$$

$$\langle f, \nu_j^N(t)\rangle = \langle f, \nu_j^N(0)\rangle + \int_0^t \langle f', \nu_j^N(u)\rangle du - \mathcal{D}_j^N(f,t) + f(0)K_j^N(t), \tag{24}$$

$$\begin{aligned}\langle \psi(\cdot,t), \eta_j^N(t)\rangle &= \langle \psi(\cdot,0), \eta_j^N(0)\rangle + \int_0^t \langle \psi_x(\cdot,u) + \psi_t(\cdot,u), \eta_j^N(u)\rangle du \\ &\quad - \mathcal{S}_j^N(\psi,t) + \int_0^t \psi(0,u) dE_j^N(u),\end{aligned} \tag{25}$$

$$\langle \varsigma, \eta_j^N(t)\rangle = \langle \varsigma, \eta_j^N(0)\rangle + \int_0^t \langle \varsigma', \eta_j^N(u)\rangle du - \mathcal{S}_j^N(\varsigma,t) + \varsigma(0)E_j^N(t). \tag{26}$$

Given $j \in \mathbb{J}$, the verification of (23) and (25) follows similarly to that given in the single class case (see Kang and Ramanan [11, proof of (2.27) and (2.28) of theorem 2.1] and Kaspi and Ramanan [12, proof of (5.4) of theorem 5.1]). Following that same line of reasoning, (24) and (26) can be verified by using the fact that the functions are compactly supported in the spatial variable and constant in the time variable in place of being compactly supported in the product space.

2.4.4. Implications of the HL Condition.

Fix $j \in \mathbb{J}$. For each $t \geq 0$, let

$$\chi_j^N(t) = \inf\{x \geq 0 : \langle 1_{[0,x]}, \eta_j^N(t)\rangle \geq Q_j^N(t)\}. \tag{27}$$

Then, for each $t \geq 0$, because there are $Q_j^N(t)$ class j customers in queue, (10) implies that $\chi_j^N(t)$ is the waiting time of the HL class j customer at time t. By definition, for each $t \geq 0$ such that $Q_j^N(t) > 0$,

$$\left\langle 1_{[0,\chi_j^N(t))}, \eta_j^N(t)\right\rangle < Q_j^N(t) \leq \left\langle 1_{[0,\chi_j^N(t)]}, \eta_j^N(t)\right\rangle. \tag{28}$$

In fact, (10) implies that for each $j \in \mathbb{J}$ and $t \geq 0$, each potential class j customer in system at time t with potential waiting time strictly less (resp. strictly greater) than $\chi_j^N(t)$ is in queue (resp. not in queue) at time t. In addition, for each $j \in \mathbb{J}$ and $t \geq 0$, there are $\left\langle 1_{[0,\chi_j^N(t)]}, \eta_j^N(t)\right\rangle - \left\langle 1_{[0,\chi_j^N(t))}, \eta_j^N(t)\right\rangle$ potential customers with potential waiting time equal to $\chi_j^N(t)$ at time t, $Q_j^N(t) - \left\langle 1_{[0,\chi_j^N(t))}, \eta_j^N(t)\right\rangle$ are in queue, and $\left\langle 1_{[0,\chi_j^N(t)]}, \eta_j^N(t)\right\rangle - Q_j^N(t)$ entered service before

they abandoned. Thus, χ_j^N is a moving boundary marking the waiting time at which class j potential customers transition from those in queue to those not in queue. This, together with the fact that χ_j^N may make a downward jump at the moment when the class j HL customer reneges, implies that, for $t \geq 0$,

$$R_j^N(t) \leq \sum_{i=-q_{0,j}^N+1}^{E_j^N(t)} \sum_{s\in[0,t]} 1_{\left\{w_{i,j}^N(s)\leq \chi_j^N(s-), \frac{dw_{j,i}^N}{dt}(s-)>0, \frac{dw_{j,i}^N}{dt}(s+)=0\right\}}, \tag{29}$$

$$R_j^N(t) \geq \sum_{i=-q_{0,j}^N+1}^{E_j^N(t)} \sum_{s\in[0,t]} 1_{\left\{w_{i,j}^N(s)< \chi_j^N(s-), \frac{dw_{j,i}^N}{dt}(s-)>0, \frac{dw_{j,i}^N}{dt}(s+)=0\right\}}. \tag{30}$$

If E_j^N has jumps of size one, η_j^N takes values in $\mathbf{M}_{D_1}[0,H_j^r)$ and (29) holds with equality. This was leveraged in analysis in Kang and Ramanan [11]. Here, the inequalities allow us to bound the reneging process both above and below by $\mathcal{S}_j^N(\psi,t)$ using appropriate functions ψ (see (56)), which is important when proving the fluid limit results in Section 4.

2.5. HL Control Policies

In Sections 2.3 and 2.4, for a deterministic initial state $y_0^N \in \mathbb{Y}^N$, we specified the dynamic equations and conditions that a $\mathbb{Y}^N$ valued process associated with a many-server queue operating under an HL control policy must satisfy. However, it is often the case that desirable policies further restrict the state of the system to take values in a subset of $\mathbb{Y}^N$. For example, nonidling control policies take values in the subset of $\mathbb{Y}^N$ given by $y \in \mathbb{Y}^N$ such that

$$N - \sum_{j=1}^J \langle 1, \nu_j\rangle = \left[N - \sum_{j=1}^J x_j\right]^+ = \left[N - \sum_{j=1}^J \langle 1,\nu_j\rangle - \sum_{j=1}^J q_j\right]^+, \tag{31}$$

where $q_j = x_j - \langle 1, \nu_j\rangle$ for all $j \in \mathbb{J}$. More generally, control policies that allow partial idleness may be desirable in overloaded systems when servers are humans (as opposed to machines) to ensure servers can rest, thereby preventing overwork and fatigue. This can be achieved by selecting a set of customer classes $\mathcal{J} \subset \mathbb{J}$ that are in some sense more important than the others and sometimes having class $j \notin \mathcal{J}$ customers wait while servers idle. Customers from class $j \in \mathcal{J}$ will only wait if all servers are busy, which is written mathematically by replacing (31) with

$$N - \sum_{j=1}^J \langle 1, \nu_j\rangle = \left[N - \sum_{j=1}^J \langle 1,\nu_j\rangle - \sum_{j\in\mathcal{J}} q_j\right]^+. \tag{32}$$

If $\mathcal{J}$ is the empty set, then (32) becomes vacuous because of (S.2).

Given a subset $\mathcal{J}$ of $\mathbb{J}$, which may be a proper subset, the empty set or the entire set, define $\mathbb{Y}_{\mathcal{J}}^N$ to be the subset of $\mathbb{Y}^N$ such that (32) holds. Note that $\mathbb{Y}_{\emptyset}^N = \mathbb{Y}^N$. Also note that for any subset $\mathcal{J}$ of $\mathbb{J}$, $\mathbb{Y}_{\mathcal{J}}^N$ is a closed subset of the Polish space $\mathbb{Y}^N$ and is therefore a Polish subspace of $\mathbb{Y}^N$. This motivates us to define an HL control policy in a way that allows for the state space to be restricted to a Polish subspace of $\mathbb{Y}^N$.

Definition 1. Given a Polish subspace $\mathbb{S}^N$ of $\mathbb{Y}^N$, an HL control policy (for the N-server queue) on $\mathbb{S}^N$ is a collection $\{\mathbb{P}_y^N\}_{y\in\mathbb{S}^N}$ of probability measures on $(\Omega,\mathcal{F})$ indexed by $\mathbb{S}^N$ such that

1. for each $y \in \mathbb{S}^N$, $\mathbb{P}_y^N\left(Y^N \in \mathbf{D}(\mathbb{S}^N), Y^N(0) = y \text{ and } Y^N \text{ satisfies (5)} - \text{(26) for } y_0^N = y\right) = 1$, and
2. for each measurable $B \subset \mathbf{D}(\mathbb{S}^N)$, the mapping $y \mapsto \mathbb{P}_y^N(B)$ from $\mathbb{S}^N$ to $[0,1]$ is Borel measurable.

Definition 1 allows one to consider random initial states satisfying some natural conditions. In particular, given a Polish subspace $\mathbb{S}^N$ of $\mathbb{Y}^N$, let

$$\begin{aligned}\mathfrak{S}^N = \{&Y^N(0) : Y^N(0) \text{is an } (\mathbb{S}^N, \mathcal{B}(\mathbb{S}^N)) \text{ valued random element defined on } (\Omega,\mathcal{F},\mathbb{P})\\ &\text{that is independent of the stochastic primitive inputs and such that}\\ &\max_{j\in\mathbb{J}} \mathbb{E}[X_j^N(0) + \langle 1, \eta_j^N(0)\rangle] < \infty, \text{ where } \mathrm{Y^N}(0)=(\alpha^{\mathrm{N}}(0), \mathrm{X^N}(0), \nu^{\mathrm{N}}(0), \eta^{\mathrm{N}}(0))\}.\end{aligned}$$

Given $Y^N(0) \in \mathfrak{S}^N$, we use the notation ς^N to denote the distribution of $Y^N(0)$; with a slight abuse of notation, we sometimes write $\varsigma^N \in \mathfrak{S}^N$. The set $\mathfrak{S}^N$ is the set of initial conditions for control policies on $\mathbb{S}^N$.

Definition 2. Given a Polish subspace $\mathbb{S}^N$ of $\mathbb{Y}^N$ and $\varsigma^N \in \mathrm{S}^N$, an HL control policy on $\mathbb{S}^N$ with initial condition ς^N is the process Y^N that for each measurable $B \subset \mathbf{D}(\mathbb{S}^N)$ satisfies

$$\mathbb{P}\left(Y^N \in B\right) = \int_{\mathbb{S}^N} \mathbb{P}_y^N(B)\varsigma^N(dy).$$

We refer to Y^N as the associated state process or simply the state process. In addition, we define $\tilde{Y}^N(0) = Y^N(0)$ and $\tilde{Y}^N(t) = (\alpha^N(t), X^N(t), \tilde{\nu}^N(t), \eta^N(t))$ for each t > 0.

2.6. Admissible HL Control Policies

In order to prove a fluid limit theorem, the entry-into-service process must satisfy some additional properties. In particular, we require that the entry-into-service process K^N is nonanticipating. Furthermore, the decision regarding which class to serve can invoke certain forms of randomness, if helpful. More specifically, given a Polish subspace $\mathbb{S}^N$ of $\mathbb{Y}^N$, $\varsigma^N \in \mathfrak{S}^N$, and an HL control policy on $\mathbb{S}^N$ with initial condition ς^N and associated state process Y^N, for each $t \geq 0$, let

$$\mathcal{G}_t^N = \sigma\left(\left\{(\tilde{Y}^N(s), 0 \leq s \leq t), \{\epsilon_i^N\}_{i=1}^{E_\Sigma^N(t)}, \{d_i^N\}_{i=1}^{D_\Sigma^N(t)}\right\}\right),$$

where $\tilde{Y}^N$ is as in Definition 2.

Definition 3. Given a Polish subspace $\mathbb{S}^N$ of $\mathbb{Y}^N$ and $\varsigma^N \in \mathfrak{S}^N$, an admissible HL control policy on $\mathbb{S}^N$ with initial condition ς^N is an HL control policy Y^N on $\mathbb{S}^N$ with initial condition ς^N (see Definition 2) such that the associated entry-into-service process K^N is $\{\mathcal{G}_t^N\}_{t\geq 0}$ adapted.

Note that $\tilde{Y}^N$ is used in the definition of the filtration to which the entry-into-service process is adapted because $\tilde{Y}^N(t)$ includes information about an arrival or a departure at time t but not about entries into service at time t. So then, for each $t > 0$, $K^N(t)$ depends the entire history $(Y^N(s), 0 < s < t)$ (because $Y^N(s) = \lim_{u \searrow s} \tilde{Y}^N(u)$) and knowledge resulting from arrivals or a departure at time t (because $\tilde{Y}^N(t)$ includes this information). Also observe that the primitive input sequences $\{\epsilon_i^N\}_{i=1}^{E_\Sigma^N(\cdot)}$ and $\{d_i^N\}_{i=1}^{D_\Sigma^N(\cdot)}$, which are i.i.d. sequences of uniform (0, 1) random variables, are included in defining the filtration to which K^N must be adapted. Therefore, at moments of an arrival event or a departure, the choice of how many customers of each class enter service, if any, can depend on the state process history and/or the independent randomness generated by these sequences.

3. A Fluid Model for the Many-Server Queue with Control

In this section, we specify a generic set of fluid model equations. These equations, given in Section 3.1, can be regarded as formal fluid analogs of the equations satisfied by any admissible HL control policy. Thus, under reasonable asymptotic conditions, one expects that fluid limit points of a fluid scaled sequence of admissible HL control policies should be fluid model solutions. Conditions for this to be true are given in Section 4 (see Theorem 1).

In Section 3.1, the fluid model equations and the definition of a fluid model solution are given (see Definition 4). We remark that the arrival function E may arise as the limit of a sequence of time-varying stochastic arrival processes and therefore, is not necessarily required to be continuous. In Section 3.2, we determine what continuity properties result for fluid model solutions upon imposing various continuity assumptions on the arrival function E and the entry-into-service function K (see Lemma 2). Also in Section 3.2 under the assumption that the arrival function is continuous, an intuitive representation for the reneging function is given (see Lemma 3). We conclude in Section 3.3 with a brief discussion of the case with constant arrival rates and the classification of the associated invariant states given Puha and Ward [19, theorem 1].

3.1. Fluid Model Solutions

The fluid model has as an input a function $E \in \mathbf{D}_\uparrow(\mathbb{R}_+^J)$, which we refer to as an arrival function. Fluid model solutions take values in the set $\mathbb{X}$ given by

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

Similarly to $\mathbb{Y}$, the set $\mathbb{X}$ endowed with the product topology is a Polish space. Given $(X, \nu, \eta) \in \mathbf{D}(\mathbb{X})$, for $j \in \mathbb{J}$, $t \geq 0$, and $x \in \mathbb{R}_+$, we set

$$F_{j,t}(x) = \langle 1_{[0,x]}, \eta_j(t)\rangle. \tag{33}$$

Also, given $(X,\nu,\eta)\in\mathbf{D}(\mathbb{X})$, for $j\in\mathbb{J}$, $t\geq 0$, and $y\in\mathbb{R}_+$, we define

$$\left(F_{j,t}\right)^{-1}(y)=\inf\{x\in\mathbb{R}_+ : F_{j,t}(x)\geq y\}, \tag{34}$$

which is taken to be infinity if $\{x\in\mathbb{R}_+ : F_{j,t}(x)\geq y\}=\emptyset$.

To define the fluid model equations, we consider a subset $\mathbf{F}$ of $\mathbf{D}(\mathbb{X})$ defined as follows: $(X,\nu,\eta)\in\mathbf{D}(\mathbb{X})$ such that for each $t\geq 0$,

$$\langle 1,\nu_j(t)\rangle\leq X_j(t)\leq \langle 1,\nu_j(t)\rangle+\langle 1,\eta_j(t)\rangle, \qquad \text{for each } j\in\mathbb{J}, \tag{35}$$

$$\sum_{j=1}^{J}\langle 1,\nu_j(t)\rangle\leq 1, \tag{36}$$

$$\int_0^t\langle h_j^s,\nu_j(u)\rangle du<\infty \quad\text{and}\quad \int_0^t\langle h_j^r,\eta_j(u)\rangle du<\infty, \qquad \text{for each } j\in\mathbb{J}. \tag{37}$$

Note that (35) and (36) are fluid analogs of (S.1) and (S.2). Condition (37) implies that the fluid analogs of the cumulative departure and potential reneging processes are finite, which holds trivially in the prelimit.

Given $(X,\nu,\eta)\in\mathbf{F}$, we define auxiliary functions B, Q, R, D, and K in $\mathbf{D}(\mathbb{R}_+^J)$ and I and $\mathbf{D}(\mathbb{R}_+)$ as follows: for each $j\in\mathbb{J}$ and $t\geq 0$,

$$B_j(t)=\langle 1,\nu_j(t)\rangle, \tag{38}$$

$$Q_j(t)=X_j(t)-B_j(t), \tag{39}$$

$$R_j(t)=\int_0^t\int_0^{Q_j(u)}h_j^r\left(\left(F_{j,u}\right)^{-1}(y)\right)dy\,du, \tag{40}$$

$$D_j(t)=\int_0^t\langle h_j^s,\nu_j(u)\rangle du, \tag{41}$$

$$K_j(t)=B_j(t)+D_j(t)-B_j(0), \tag{42}$$

$$I(t)=1-\sum_{j=1}^{J}B_j(t). \tag{43}$$

Then B, Q, R, D, K, and I are fluid analogs of the busy server, the queue length, the reneging, the departure, the entry-into-service, and the idleness processes, respectively. For $(X,\nu,\eta)\in\mathbf{F}$, $0\leq I(t)\leq 1$ for all $t\geq 0$ because of (36), (38), and (43), and $0\leq Q_j(t)\leq\langle 1,\eta_j(t)\rangle$ for all $t\geq 0$ and $j\in\mathbb{J}$ because of (35) and (39). Furthermore, because of (37), $D_j(t)$ and $R_j(t)$ are finite for all $t\geq 0$ and $j\in\mathbb{J}$ and therefore, $K_j(t)$ is finite for all $t\geq 0$ and $j\in\mathbb{J}$.

Next, we introduce some additional properties and equations that should be satisfied by $(X,\nu,\eta)\in\mathbf{F}$ arising as fluid limit points. Because $K\in\mathbf{D}(\mathbb{R}_+^J)$ is to be regarded as the fluid analog of the entry-into-service process, it should satisfy that for each $j\in\mathbb{J}$,

$$K_j \text{ is nondecreaing.} \tag{44}$$

Further the fluid analog of the balance Equation (17) should hold. In particular, for all $j\in\mathbb{J}$ and $t\geq 0$,

$$X_j(t)=X_j(0)+E_j(t)-R_j(t)-D_j(t). \tag{45}$$

In addition, ν and η should satisfy certain integral equations. In particular, for all $j\in\mathbb{J}$, $f\in\mathbf{C}_b(\mathbb{R}_+)$, and $t\geq 0$,

$$\langle f,\nu_j(t)\rangle=\left\langle f(\cdot+t)\frac{1-G^s(\cdot+t)}{1-G^s(\cdot)},\nu_j(0)\right\rangle+\int_0^t f(t-u)(1-G_j^s(t-u))dK_j(u), \tag{46}$$

$$\langle f,\eta_j(t)\rangle=\left\langle f(\cdot+t)\frac{1-G^r(\cdot+t)}{1-G^r(\cdot)},\eta_j(0)\right\rangle+\int_0^t f(t-u)(1-G_j^r(t-u))dE_j(u). \tag{47}$$

Definition 4. Let E be an arrival function. A fluid model solution for E is $(X,\nu,\eta)\in\mathbf{F}$ that satisfies (44) and the fluid model Equations (45)–(47).

Remark 1. We chose the more intuitive Equations (46) and (47) in the fluid model in place of the fluid analogs of (23) and (25) in Lemma 1, which are given as follows: for all $j\in\mathbb{J}$, $\varphi\in\mathbf{C}_c^{1,1}([0,H_j^s)\times\mathbb{R}_+)$, $\psi\in\mathbf{C}_c^{1,1}([0,H_j^r)\times\mathbb{R}_+)$,

and $t \geq 0$,

$$\begin{aligned}\langle \varphi(\cdot,t), \nu_j(t)\rangle &= \langle \varphi(\cdot,0), \nu_j(0)\rangle + \int_0^t \langle \varphi_x(\cdot,u) + \varphi_t(\cdot,u), \nu_j(u)\rangle du \\ &\quad - \int_0^t \langle h_j^s(\cdot)\varphi(\cdot,u), \nu_j(u)\rangle du + \int_0^t \varphi(0,u) dK_j(u),\end{aligned} \tag{48}$$

$$\begin{aligned}\langle \psi(\cdot,t), \eta_j(t)\rangle &= \langle \psi(\cdot,0), \eta_j(0)\rangle + \int_0^t \langle \psi_x(\cdot,u) + \psi_t(\cdot,u), \eta_j(u)\rangle du \\ &\quad - \int_0^t \langle h_i^r(\cdot)\psi(\cdot,u), \eta_j(u)\rangle du + \int_0^t \psi(0,u) dE_j(u).\end{aligned} \tag{49}$$

By Kaspi and Ramanan [12, theorem 4.1], if $(\nu,\eta) \in \mathbf{D}(\times_{j=1}^J \mathbf{M}[0,H_j^s) \times \times_{j=1}^J \mathbf{M}[0,H_j^r))$ satisfies (37) and $K \in \mathbf{D}(\mathbb{R}_+^J)$ satisfies $K_j(0) = 0$ for all $j \in \mathbb{J}$ and (44), then we have that (46) and (47) hold if and only if (48) and (49) hold. Therefore, our choice is not restrictive and consistent with the prior works Atar et al. [4], Kang and Ramanan [11], and Kaspi and Ramanan [12]. In addition, as in Kang and Ramanan [11, corollary 4.2], if $(\nu,\eta) \in \mathbf{D}(\times_{j=1}^J \mathbf{M}[0,H_j^s) \times\times_{j=1}^J \mathbf{M}[0,H_j^r))$ satisfies (37) and $K \in \mathbf{D}(\mathbb{R}_+^J)$ satisfies $K_j(0) = 0$ for all $j \in \mathbb{J}$ and (44), then (46) and (47) hold for each $f \in \mathbf{C}_b(\mathbb{R}_+)$ if and only if (46) and (47) hold for each bounded Borel measurable function $f : \mathbb{R}_+ \to \mathbb{R}$.

3.2. Continuity Properties of Fluid Model Solutions

This lemma is a modest generalization of Kang and Ramanan [11, corollary 3.7] and Kaspi and Ramanan [12, lemmas 5.18 and 7.3].

Lemma 2. *Suppose that E is an arrival function and (X,ν,η) is a fluid model solution for arrival function E. Then the following hold.*

i. *If the arrival function E and the auxiliary function K are continuous, then (X,ν,η) and each of the auxiliary functions are continuous.*

ii. *If, for each $j \in \mathbb{J}$, E_j and K_j are absolutely continuous, then X_j and all of the coordinates of the auxiliary functions are absolutely continuous for each $j \in \mathbb{J}$.*

iii. *If, K (resp. E) is continuous and for each $j \in \mathbb{J}$, $\nu_j(0)$ (resp. $\eta_j(0)$) has no atoms, then $\nu_j(t)$ (resp. $\eta_j(t)$) has no atoms for each $t \geq 0$ and $j \in \mathbb{J}$.*

iv. *If, for each $j \in \mathbb{J}$, K_j (resp. E_j) is absolutely continuous and $\nu_j(0)$ (resp. $\eta_j(0)$) is absolutely continuous with respect to Lebesgue measure, then $\nu_j(t)$ (resp. $\eta_j(t)$) is absolutely continuous with respect to Lebesgue measure for each $t \geq 0$ and $j \in \mathbb{J}$.*

Proof. First we verify (i). To begin, fix $j \in \mathbb{J}$. Continuity of $t \mapsto \nu_j(t)$ (resp. $t \mapsto \eta_j(t)$) with respect to the topology of weak convergence follows from (46) (resp. (47)), bounded convergence, and continuity of K (resp. E). By (41) and (40), D_j and R_j are absolutely continuous and therefore continuous. Then, by (45) and (42), X_j and B_j are continuous. Hence, by (39), Q_j is continuous. Because j was arbitrary, I is continuous by (43). The proof of (ii) follows a similar line of reasoning. Hence (i) and (ii) hold. To verify (iii), fix $j \in \mathbb{J}$ and $x \in \mathbb{R}_+$ and substitute $f = 1_{\{x\}}$ into (46) and (47), which is valid because of Remark 1, and observe that the right-hand side is zero for all $t \geq 0$. Since j and x were arbitrary, (iii) holds. Statement (iv) can be verified by adapting the argument used to proof Kaspi and Ramanan [12, lemma 5.18]. □

A simplified fluid model is discussed in the tutorial paper Puha and Ward [19], where it is assumed that E is continuous and $\eta_j(0)$ has no atoms for each $j \in \mathbb{J}$. An advantage to making this assumption is that one finds an alternative, perhaps more intuitive, representation for the reneging function, which is presented in the next lemma. For this, given a fluid model solution (X,ν,η) for an arrival function E, it will be convenient to have a notation for the fluid analog χ of the process χ^N, which is a function taking values in $\mathbb{R}_+^J$ given as follows: for $j \in \mathbb{J}$ and $t \geq 0$, let

$$\chi_j(t) = \inf\{x \in \mathbb{R}_+ : \langle 1_{[0,x]}, \eta_j(t)\rangle \geq Q_j(t)\}. \tag{50}$$

Lemma 3. *Suppose that E is a continuous arrival function and (X,ν,η) is a fluid model solution for arrival function E such that $\eta_j(0)$ has no atoms for each $j \in \mathbb{J}$. Then, for each $j \in \mathbb{J}$ and $t \geq 0$,*

$$R_j(t) = \int_0^t \int_0^{H_j^r} h_j^r(x) 1_{\{\langle 1_{[0,x]}, \eta_j(u)\rangle < Q_j(u)\}} \eta_j(u)(dx)du. \tag{51}$$

Proof. Fix $j \in \mathbb{J}$ and $t \geq 0$. By definition of χ_j,

$$\int_0^t \int_0^{H_j^r} h_j^r(x) 1_{\{\langle 1_{[0,x]}, \eta_j(u)\rangle < Q_j(u)\}} \eta_j(u)(dx)du = \int_0^t \int_{[0,\chi_j(u))} h_j^r(x) F_{j,u}(dx)du. \tag{52}$$

Then, given $0 \leq u \leq t$, perform the change of variables $z = \left(F_{j,u}\right)^{-1}(y)$ and use the fact that because of Lemma 2 (iii), $\eta_j(u)$ does not have any atoms to find

$$\int_{[0,\chi_j(u))} h_j^r(x) F_{j,u}(dx) = \int_{[0,Q_j(u))} h_j^r\left(\left(F_{j,u}\right)^{-1}(y)\right)dy.$$

This together with (52) and (40) implies (51). □

3.3. Discussion of Constant Arrival Rates and Invariant States

In Puha and Ward [19], the authors characterize the invariant states, or fixed points of the fluid model equations in the case where the arrival rate to each class is constant (i.e., $E_j(t) = \lambda_j t$ for all $t \geq 0$ and $j \in \mathbb{J}$ for some $\lambda \in (0,\infty)^J$). Given $\lambda \in (0,\infty)^J$, we define the function $\Lambda_j(t) = \lambda_j t$ for all $t \geq 0$ and $j \in \mathbb{J}$.

Definition 5. A tuple $(X^*, \nu^*, \eta^*) \in \mathbf{F}$ is an invariant state for the arrival function E 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 for the arrival function E. (i.e., satisfies Definition 4). Given $(X^*, \nu^*, \eta^*) \in \mathbf{F}$, for convenience we define $B_j^* = \langle 1, \nu_j^* \rangle$ and $Q_j^* = X_j^* - B_j^*$ for each $j \in \mathbb{J}$.

Proposition 1 (Theorem 1 in Puha and Ward [19] Restated for the Reader's Convenience). *Suppose that $\lambda \in (0,\infty)^J$ and that for each $j \in \mathbb{J}$, $m_j^s = \int_0^{H_j^s} 1 - G_j^s(x)dx < \infty$, $m_j^r = \int_0^{H_j^r} 1 - G_j^r(x)dx < \infty$ and G_j^r is strictly increasing with inverse $(G_j^r)^{-1}$, where by convention $(G_j^r)^{-1}(1) = H_j^r$. Define*

$$\mathbb{B}(\lambda) = \left\{ b \in \mathbb{R}_+^J : b_j \leq \rho_j := \lambda_j m_j^s \text{ for all } j \in \mathbb{J} \text{ and } \sum_{j=1}^J b_j \leq 1 \right\}. \tag{53}$$

For $b \in \mathbb{B}(\lambda)$ and $j \in \mathbb{J}$, define $q(b) \in \mathbb{R}_+^J$ such that

$$q_j(b) = \lambda_j \int_0^{(G_j^r)^{-1}(1-b_j/\rho_j)} (1 - G_j^r(x))dx, \tag{54}$$

and let

i. $\eta_j^*(dx) = \lambda_j(1 - G_j^r(x))dx$ *for each* $x \in \mathbb{R}_+$,

ii. $\nu_j^*(dx) = \frac{b_j}{m_j^s}(1 - G_j^s(x))dx$ *for each* $x \in \mathbb{R}_+$, *and*

iii. $X_j^* = q_j(b_j) + b_j$.

iv. *Then (X^*, ν^*, η^*) is an invariant state for the arrival function Λ with $B^* = b$ and $Q^* = q(b)$. Conversely, if (X^*, ν^*, η^*) is an invariant state for the arrival function Λ, then $B^* \in \mathbb{B}(\lambda)$ and (X^*, ν^*, η^*) satisfies (i)–(iii) for $b = B^*$*[2].

For $\lambda \in (0,\infty)^J$, Proposition 1 implies that the invariant states for an arrival function Λ are in one-to-one correspondence with the set $\mathbb{B}(\lambda)$. In particular, $b \in \mathbb{B}(\lambda)$ captures the long-run average fraction of the collective server effort that must be provided to each class by the control policy for it to be possible to reach an equilibrium in the invariant state associated with b. There are $b \in \mathbb{B}(\lambda)$ such that $0 < b_j < \lambda_j m_j^s$ for more than one $j \in \mathbb{J}$, and, for such a b, multiple classes must be partially served.

4. Fluid Limit Points for Many-Server Queues with Control

]In this section, we consider a sequence of multiclass many-server queues indexed by N, the number of servers. For each $N \in \mathbb{N}$, there is an arrival process E^N, a Polish subspace $\mathbb{S}^N$ of $\mathbb{Y}^N$, $\varsigma^N \in \mathfrak{S}^N$, and an admissible HL control policy on $\mathbb{S}^N$ with initial condition ς^N. Then, for each $N \in \mathbb{N}$, Y^N denotes the associated state process (see Definition 3). We define the fluid scaled state processes for the Nth system as follows. To begin, let $\bar{\alpha}^N = \alpha^N$. Also, for $H^N = E^N$, X^N, ν^N, η^N, B^N, Q^N, R^N, D^N, D_Σ^N, K^N, and I^N, let $\bar{H}^N = H^N/N$. Then $\bar{Y}^N = (\bar{\alpha}^N, \bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)$ for all $N \in \mathbb{N}$. The main result proved in this section is a fluid limit theorem that gives conditions under which limit points of the sequence $\{\bar{Y}^N\}_{N\in\mathbb{N}}$ are fluid model solutions (see Theorem 1). To state Theorem 1, the conditions that $\{\bar{Y}^N\}_{N\in\mathbb{N}}$ must satisfy need to be given.

Assumption 1. *E is an arrival process (which may be a random element defined on $(\Omega,\mathcal{F},\mathbb{P})$ taking values in $\mathbf{D}\uparrow(\mathbb{R}_+^J)$) such that for each $j\in\mathbb{J}$,*

1. $\lim_{N\to\infty}\bar{E}_j^N = E_j$ *almost surely;*
2. $\lim_{N\to\infty}\mathbb{E}\left[\bar{E}_j^N(t)\right] = \mathbb{E}\left[E_j(t)\right] < \infty$ *for all* $t\geq 0$.

We remark that Assumption 1 allows for E to possibly have jump discontinuities. Continuity is not necessary to verify tightness. This is demonstrated in Section 4.2 (see Theorem 2). We also require some regularity on the entry-into-service process, which is stated next as Assumption 2.

Assumption 2. *Either $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$ satisfies (K.2) of Kurtz' criteria (stated in Section A.1 in the appendix) for each $j\in\mathbb{J}$ or for all $N\in\mathbb{N}$ and $0\leq s<t<\infty$,*

$$\max_{j\in\mathbb{J}} K_j^N(t) - K_j^N(s) \leq E_\Sigma^N(t) + D_\Sigma^N(t) - E_\Sigma^N(s) - D_\Sigma^N(s). \tag{55}$$

The inequalities (55) imply that the number of customers of a particular class that enter service at any moment is bounded above by the total number of customers that arrived and departed at that moment. It is satisfied by many natural admissible HL control policies. When it holds, there is no need to verify (K.2) of Kurtz' criteria for $\{\bar{K}^N\}_{N\in\mathbb{N}}$ a priori in order to apply the results here. Instead, in the proof of Lemma 8 in the case where (55) is assumed to hold, we use (55) together with the other conditions to verify that $\{\bar{K}^N\}_{N\in\mathbb{N}}$ is tight. This might seem reasonable since (55) implies that for each $j\in\mathbb{J}$, the oscillations of $\bar{K}_j^N$ are controlled by the oscillations of $\bar{E}_\Sigma^N + \bar{D}_\Sigma^N$.

Finally, the next assumption provides control over the asymptotic behavior of the sequence of initial conditions.

Assumption 3. *(X^0,ν^0,η^0) is a random element defined on $(\Omega,\mathcal{F},\mathbb{P})$ taking values in $\mathbb{X}$ such that for each $j\in\mathbb{J}$,*

1. $\lim_{N\to\infty}\bar{X}_j^N(0) = X_j^0$ *almost surely;*
2. $\lim_{N\to\infty}\mathbb{E}[\bar{X}_j^N(0)] = \mathbb{E}[X_j^0] < \infty$;
3. $\bar{\nu}_j^N(0)\xrightarrow{w}\nu_j^0$ *as* $N\to\infty$ *almost surely;*
4. $\bar{\eta}_j^N(0)\xrightarrow{w}\eta_j^0$ *as* $N\to\infty$ *almost surely;*
5. $\lim_{N\to\infty}\mathbb{E}\left[\langle 1,\bar{\eta}_j^N(0)\rangle\right] = \mathbb{E}[\langle 1,\eta_j^0\rangle] < \infty$.

Assumptions 1, 2, and 3 are needed to prove the main tightness result, Theorem 2, which is stated in Section 4.2 after developing the requisite background. These assumptions say nothing about the convergence of $\langle h_j^s,\bar{\nu}_j^N(0)\rangle$ or of $\langle h_j^r,\bar{\eta}_j^N(0)\rangle$ as $N\to\infty$, for $j\in\mathbb{J}$. It turns out that this is not needed to prove tightness. However, additional conditions must be satisfied in order to uniquely characterize limit points. This can be achieved by imposing some mild conditions on the hazard functions as follows.

Assumption 4. *For each $j\in\mathbb{J}$, there exists $L_j^s < H_j^s$ such that h_j^s is either bounded or lower-semicontinuous on (L_j^s,H_j^s). Likewise, for each $j\in\mathbb{J}$, there exists $L_j^r < H_j^r$ such that h_j^r is either bounded or lower-semicontinuous on (L_j^r,H_j^r).*

Furthermore, in order to uniquely characterize limit points, and in particular to characterize the limit of the reneging process, the following assumption is needed. For this, we recall that a measure does not charge points if it assigns zero mass to all singletons.

Assumption 5. *Assumptions 1, 2, and 3 hold and the following hold:*

1. *E_j is continuous almost surely for each $j\in\mathbb{J}$;*
2. *η_j^0 does not charge points almost surely, for each $j\in\mathbb{J}$;*
3. *either $\{\bar{K}^N\}_{N\in\mathbb{N}}$ is $\mathbf{C}$-tight, or (55) holds.*

Part 1 of Assumption 5 holds in many cases arising in applications. Part 2 in Assumption 5 is a condition on η^0 that is preserved in time by the fluid model dynamics in the presence of Part 1 in Assumption 5 (see Lemma 2(iii)). We show that this holds for fluid limit points in the proof of in Lemma 10(6), which generalizes Kang and Ramanan [11, lemma 7.3] to the present setting.

Theorem 1. *Suppose that Assumptions 4 and 5 hold. Then $\{(\bar{X}^N,\bar{\nu}^N,\bar{\eta}^N)\}_{N\in\mathbb{N}}$ is tight. If (X,ν,η) is a distributional limit point of $\{(\bar{X}^N,\bar{\nu}^N,\bar{\eta}^N)\}_{N\in\mathbb{N}}$,* then *$(X,\nu,\eta)$ is, almost surely, a fluid model solution for E with $(X(0),\nu(0),\eta(0)) = (X^0,\nu^0,\eta^0)$. Furthermore, (X,ν,η) and the auxiliary functions are continuous and $\eta_j(t)$ does not charge points for all $t\geq 0$ and $j\in\mathbb{J}$, almost surely.*

Theorem 1 is proved in Section 4.3 after establishing some preliminaries in Section 4.1 and a more complete tightness result in Section 4.2. It follows closely the general outline of the proof of Kang and Ramanan [11, theorem 7.1], which considers the single class, nonidling case where the arrival process has jumps of size one. Here, we treat the multiclass case with batch arrivals and more general admissible HL control policies that may idle. Our development emphasizes the nuances that must be addressed in this case.

We also remark that a foundation for allowing the limiting arrival process to have jumps is developed in Kang and Ramanan [11]. However, in Kang and Ramanan [11], E^N is assumed to have jumps of size one, which implies that Part 1 of Assumption 5 holds. More importantly, the alternative representation of the reneging process compensator given in Kang and Ramanan [11, proposition 5.5] relies on the fact that E^N is assumed to have jumps of size one. Here, we have relaxed this by allowing batch arrivals. To account for this, we provide upper and lower bounds on the prelimit reneging process (see (56)). The characterization of the limit is executed by analyzing the martingale compensators for these upper and lower bounds. Assumption 5 is used to argue that the fluid limits of the martingale compensators for these upper and lower bounds agree. The characterization of the limiting reneging process is shown to follow from this in the proof of Lemma 10(7).

Theorem 1 does not imply convergence; but rather, it characterizes limit points as fluid model solutions. A lack of uniqueness of fluid model solutions due to the absence of a policy-specific equation or equations may result in different distributional limits along different subsequences. If one has a specific collection of admissible HL control policies that they wish to analyze, then Theorem 1 can be used as a step toward proving a fluid limit theorem. A program for this is outlined in Section 4.4 immediately following the proof of Theorem 1.

4.1. The Collections of Marked Point Processes

In view of (23) and (25), weak convergence of the sequence of rescaled state descriptors involves weak convergence of the marked point processes $\{\bar{\mathcal{D}}_j^N(\varphi,\cdot)\}_{N\in\mathbb{N}}$ and $\{\bar{\mathcal{S}}_j^N(\psi,\cdot)\}_{N\in\mathbb{N}}$ for each $j\in\mathbb{J}$, $\varphi\in\mathbf{C}_c^{1,1}([0,H_j^s)\times\mathbb{R}_+)$ and $\psi\in\mathbf{C}_c^{1,1}([0,H_j^r)\times\mathbb{R}_+)$. In this section, we establish some facts about these collections of marked point processes. This generalizes and modestly restructures the development in Kang and Ramanan [11] and Kaspi and Ramanan [12], which facilitates applying it in the present multiclass setting with batch arrivals. Although we do not find it necessary to write out the proofs of most results, we find value in stating the relevant lemmas, which are slightly more general in some cases, and commenting on their proofs.

4.1.1. Families of Martingales. The analysis will make use of certain families of martingales. In order to introduce these families, we must define an appropriate filtration. For this, recall that for each $N\in\mathbb{N}$, $j\in\mathbb{J}$, and $i=-X_j^N(0)+1,\ldots,0,1,2,\ldots$, $k_{j,i}^N$ is the time at which the i^{th} class j customer enters service, which is taken to be infinity if that customer reneges before entering service. For bookkeeping purposes, we label each of the N-servers with a distinct index from the set $\{1,2,\ldots,N\}$. For each $N\in\mathbb{N}$, $j\in\mathbb{J}$, $t\geq 0$, and $i=-X_j^N(0)+1,\ldots,0,1,2,\ldots$, we let $s_{j,i}^N(t)\in\{0,1,2,\ldots,N\}$ be the index of the server that processed or is currently processing the work associated with class j customer i if $k_{j,i}^N\leq t$ and is zero otherwise. For each $N\in\mathbb{N}$ and $t\geq 0$, let

$$\begin{aligned}\tilde{\mathcal{F}}_t^N=\sigma\Big(&\Big\{Y^N(0),\Big(\alpha^N(u),0\leq u\leq t\Big),\Big\{\Big\{\Big(w_{j,i}^N(u),0\leq u\leq t\Big)\Big\}_{i=\langle 1,\eta_j^N(0)\rangle+1}^{\infty}\Big\}_{j\in\mathbb{J}},\\&\Big\{\Big\{\Big(a_{j,i}^N(u),0\leq u\leq t\Big)\Big\}_{i=-X_j^N(0)+1}^{\infty}\Big\}_{j\in\mathbb{J}},\Big\{\Big\{\Big(s_{j,i}^N(u),0\leq u\leq t\Big)\Big\}_{i=-X_j^N(0)+1}^{\infty}\Big\}_{j\in\mathbb{J}},\\&\{\epsilon_i^N\}_{i=1}^{E_\Sigma^N(t)},\{d_i^N\}_{i=1}^{D_\Sigma^N(t)}\Big\}\Big),\end{aligned}$$

and, for each $N\in\mathbb{N}$, let $\{\mathcal{F}_t^N\}_{t\geq 0}$ be the associated right-continuous filtration, completed with respect to $\mathbb{P}$. For each $N\in\mathbb{N}$, Y^N is an $\{\mathcal{F}_t^N\}_{t\geq 0}$ adapted process and $\mathcal{G}_t^N\subseteq\mathcal{F}_t^N$ for all $t\geq 0$. If $J=1$ and the control policy is the nonidling policy as in Kang and Ramanan [11], then the i.i.d. sequences $\{\epsilon_i^N\}_{i=1}^\infty$ and $\{d_i^N\}_{i=1}^\infty$ of uniform (0, 1) random variables are not used to determine the entry-into-service process K^N. In this case, the filtration considered here is equal to the filtration defined in Kang and Ramanan [11, section 2.2.4], augmented with the sequences $\{\epsilon_i^N\}_{i=1}^\infty$ and $\{d_i^N\}_{i=1}^\infty$.

For each $j\in\mathbb{J}$, bounded measurable function $\varphi:[0,H_j^s)\times\mathbb{R}_+\to\mathbb{R}$, bounded measurable function $\psi:[0,H_j^r)\times\mathbb{R}_+\to\mathbb{R}$, and $t\geq 0$, let

$$\mathcal{A}_{s,j}^N(\varphi,t)=\int_0^t\langle\varphi(\cdot,u)h_j^s(\cdot),\nu_j^N(u)\rangle du\quad\text{and}\quad\mathcal{M}_{s,j}^N(\varphi,t)=\mathcal{D}_j^N(\varphi,t)-\mathcal{A}_{s,j}^N(\varphi,t),$$

$$\mathcal{A}_{r,j}^N(\psi,t)=\int_0^t\langle\psi(\cdot,u)h_j^r(\cdot),\eta_j^N(u)\rangle du\quad\text{and}\quad\mathcal{M}_{r,j}^N(\psi,t)=\mathcal{S}_j^N(\psi,t)-\mathcal{A}_{r,j}^N(\psi,t).$$

In addition, for $j \in \mathbb{J}$, $w \in [0, H_j^r)$, and $t \geq 0$, let

$$\theta_j^N(w,t) = 1_{(w,\infty)}(\chi_j^N(t-)) = 1_{[0,\chi_j^N(t-))}(w),$$
$$\Theta_j^N(w,t) = 1_{[w,\infty)}(\chi_j^N(t-)) = 1_{[0,\chi_j^N(t-)]}(w).$$

For each $j \in \mathbb{J}$, θ_j^N and Θ_j^N are almost surely bounded, measurable, real-valued functions on $[0, H_j^s) \times \mathbb{R}_+$. Thus, for each $j \in \mathbb{J}$ and $t \geq 0$, $\mathcal{A}_{r,j}^N(\theta_j^N, t)$ and $\mathcal{A}_{r,j}^N(\Theta_j^N, t)$ are well defined. Furthermore, by (29) and (30), for each $j \in \mathbb{J}$ and $t \geq 0$,

$$\mathcal{S}_j^N(\theta_j^N, t) \leq R_j^N(t) \leq \mathcal{S}_j^N(\Theta_j^N, t). \tag{56}$$

By adapting the arguments used to prove Kaspi and Ramanan [12, corollary 5.5], part 1 of Kang and Ramanan [11, proposition 5.1], and [11, lemma 5.4] to the present multiclass setting with batch arrivals, the following lemma holds. This relies on the independence conditions satisfied by the stochastic primitive inputs and the initial conditions, and the requirement that K^N is $\{\mathcal{G}_t^N\}_{t\geq 0}$ adapted in Definition 3.

Lemma 4. *Let $N \in \mathbb{N}$. For each $j \in \mathbb{J}$, bounded measurable function $\varphi : [0, H_j^s) \times \mathbb{R}_+ \to \mathbb{R}$ such that $t \to \varphi(a_{j,i}^N(t), t)$ is left continuous on $[0,\infty)$ for each $i \in \{-X_j^N(0)+1, \ldots, 0\} \cup \mathbb{N}$, and bounded measurable function $\psi : [0, H_j^r) \times \mathbb{R}_+ \to \mathbb{R}$ such that $t \to \psi(w_{j,i}^N(t), t)$ is left continuous on $[0,\infty)$ for each $i \in \{-\langle 1, \eta_j^N(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$, the processes $\mathcal{A}_{s,j}^N(\varphi, \cdot)$ and $\mathcal{A}_{r,j}^N(\psi, \cdot)$ are the $\{\mathcal{F}_t^N\}_{t\geq 0}$-compensators of $\mathcal{D}_j^N(\varphi, \cdot)$ and $\mathcal{S}_j^N(\psi, \cdot)$ respectively. Also, $\mathcal{A}_{r,j}(\Theta_j^N, \cdot)$ is the $\{\mathcal{F}_t^N\}_{t\geq 0}$-compensator of $\mathcal{S}_j^N(\Theta_j^N, \cdot)$. In particular, the processes $\mathcal{M}_{s,j}^N(\varphi, \cdot)$, $\mathcal{M}_{r,j}^N(\psi, \cdot)$, and $\mathcal{M}_{r,j}^N(\Theta_j^N, \cdot)$ are local $\{\mathcal{F}_t^N\}_{t\geq 0}$-martingales.*

Remark 2. For each $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $i \in \{-\langle 1, \eta_j^N(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$, $\theta_j^N(w_{j,i}^N(\cdot), \cdot)$ is left continuous on $[0,\infty)$, while $\Theta_j^N(w_{j,i}^N(\cdot), \cdot)$ is not left continuous on $[0,\infty)$. Hence, the martingale property involving Θ_j^N is stated separately in Lemma 4. For the single class case, an approximation argument is given in the proof Kang and Ramanan [11, lemma 5.4] to address the lack of left continuity, and that argument extends to the multiclass case as well.

For each $j \in \mathbb{J}$, bounded measurable function $\varphi : [0, H_j^s) \times \mathbb{R}_+ \to \mathbb{R}$, bounded measurable function $\psi : [0, H_j^r) \times \mathbb{R}_+ \to \mathbb{R}$, and $H^N = \mathcal{D}_j^N(\varphi, \cdot), \mathcal{A}_{s,j}^N(\varphi, \cdot), \mathcal{M}_{s,j}^N(\varphi, \cdot), \mathcal{S}_j^N(\psi, \cdot), \mathcal{A}_{r,j}^N(\psi, \cdot)$, and $\mathcal{M}_{r,j}^N(\psi, \cdot)$, let $\bar{H}^N = H^N/N$. Each of the local martingales given in Lemma 4 has a well-defined quadratic variation process, which tends to zero under fluid scaling and the asymptotic assumptions specified in the next lemma. To state this, given a local martingale L^N, we let $\bar{L}^N = L^N/N$ and let $\langle \bar{L}^N \rangle(\cdot)$ denote quadratic variation process associated with $\bar{L}^N$, if such a process is well-defined.

Lemma 5. *Suppose that Assumptions 1 and 3 hold. For each $t \geq 0$, $j \in \mathbb{J}$, bounded measurable function $\varphi : [0, H_j^s) \times \mathbb{R}_+ \to \mathbb{R}$ such that $u \mapsto \varphi(a_{j,i}^N(u), u)$ is left continuous on $[0,\infty)$ for each $i \in \{-X_j^N(0)+1, \ldots, 0\} \cup \mathbb{N}$, and bounded measurable function $\psi : [0, H_j^r) \times \mathbb{R}_+ \to \mathbb{R}$ such that $u \mapsto \psi(w_{j,i}^N(u), u)$ is left continuous on $[0,\infty)$ for each $i \in \{-\langle 1, \eta_j^N(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$,*

$$\limsup_{N\to\infty} \mathbb{E}\left[\bar{H}^N(t)\right] < \infty, \tag{57}$$

for $H^N(\cdot) = \mathcal{D}_j^N(\varphi, \cdot), \mathcal{A}_{s,j}^N(\varphi, \cdot), \mathcal{S}_j^N(\psi, \cdot), \mathcal{A}_{r,j}^N(\psi, \cdot), \mathcal{S}_j^N(\Theta_j^N, \cdot)$, and $\mathcal{A}_{r,j}^N(\Theta_j^N, \cdot)$. Consequently,

$$\limsup_{N\to\infty} \mathbb{E}\left[\bar{R}_j^N(t)\right] < \infty. \tag{58}$$

Furthermore, for $\bar{L}^N(\cdot) = \bar{\mathcal{M}}_{s,j}^N(\varphi, \cdot), \bar{\mathcal{M}}_{r,j}^N(\psi, \cdot)$ and $\bar{\mathcal{M}}_{r,j}^N(\Theta_j^N, \cdot)$,

$$\lim_{N\to\infty} \mathbb{E}\left[\langle \bar{L}^N \rangle(t)\right] = 0, \qquad \textit{so that} \qquad \bar{L}^N \Rightarrow \mathbf{0}, \text{ as } N \to \infty. \tag{59}$$

Proof. Fix $t \geq 0$ and $j \in \mathbb{J}$. As in the proof of Kaspi and Ramanan [12, part 1 of lemma 5.6], (57) follows from (21), (22), Lemma 4 (which implies that for each $N \in \mathbb{N}$, $\mathbb{E}[\bar{\mathcal{D}}_j^N(\varphi, t)] = \mathbb{E}[\bar{\mathcal{A}}_{s,j}^N(\varphi, t)]$ for each bounded measurable

function $\varphi : [0, H_j^s) \times \mathbb{R}_+ \to \mathbb{R}$ such that $u \mapsto \varphi(a_{j,i}^N(u), u)$ is left continuous on $[0, \infty)$ for each $i \in \{-X_j^N(0) + 1, \ldots, 0\} \cup \mathbb{N}$, $\mathbb{E}[\bar{\mathcal{S}}_j^N(\psi, t)] = \mathbb{E}[\bar{\mathcal{A}}_{r,j}^N(\psi, t)]$ for each bounded measurable function $\psi : [0, H_j^r) \times \mathbb{R}_+ \to \mathbb{R}$ such that $u \mapsto \psi(w_{j,i}^N(u), u)$ is left continuous on $[0, \infty)$ for each $i \in \{-\langle 1, \eta_j^N(0)\rangle + 1, \ldots, 0\} \cup \mathbb{N}$, and $\mathbb{E}[\bar{R}_j^N(t)] \le \mathbb{E}[\bar{\mathcal{A}}_{r,j}^N(\Theta_j^N, t)]$), and Assumptions 1 and 3. Then (59) follows from (57) as in the proof of Kaspi and Ramanan [12, lemma 5.9]. □

Next we develop alternative representations for the compensators of $\mathcal{S}_j^N(\theta_j^N, \cdot)$ and $\mathcal{S}_j^N(\Theta_j^N, \cdot)$ for each $j \in \mathbb{J}$ and $N \in \mathbb{N}$. This is similar to the development in Kang and Ramanan [11] and Kaspi and Ramanan [12], but we require a modest generalization to account for batch arrivals. For $N \in \mathbb{N}$, $j \in \mathbb{J}$, $t \ge 0$, and $x \in [0, H_j^r)$, let

$$F_{j,t}^N(x) = \langle 1_{[0,x]}, \eta_j^N(t)\rangle.$$

For $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $t \ge 0$, let

$$\tilde{\chi}_j^N(t) = \inf\left\{x \in [0, H_j^r) : F_{j,t}^N(x) \ge \left\langle 1_{[0,\chi_j^N(t-))}, \eta_j^N(t)\right\rangle\right\}.$$

Then, for $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $t \ge 0$, either $\tilde{\chi}_j^N(t) = \chi_j^N(t-) = 0$ or $0 \le \tilde{\chi}_j^N(t) < \chi_j^N(t-)$ and $\left\langle 1_{(\tilde{\chi}_j^N(t), \chi_j^N(t-))}, \eta_j^N(t)\right\rangle = 0$. Hence, for $N \in \mathbb{N}$, $j \in \mathbb{J}$, $t \ge 0$, and $w \in \mathbb{R}_+$, $\theta_j^N(w) = 1_{[0,\tilde{\chi}_j^N(t)]}(w)$. Therefore, for $N \in \mathbb{N}$, $j \in \mathbb{J}$, $t \ge 0$,

$$\mathcal{A}_{r,j}^N(\theta_j^N, t) = \int_0^t \left\langle 1_{[0,\tilde{\chi}_j^N(u)]}(\cdot) h_j^r(\cdot), \eta_j^N(u)\right\rangle du. \tag{60}$$

Lemma 6 (Alternative Compensator Process Representation). *For each $N \in \mathbb{N}$, $j \in \mathbb{J}$, $t \ge 0$, and $x \in [0, H_j^r)$,*

$$\langle 1_{[0,x]} h_j^r, \eta_j^N(t)\rangle = \int_0^{F_{j,t}^N(x)} h_j^r\left(\left(F_{j,t}^N\right)^{-1}(y)\right) dy. \tag{61}$$

In particular, for each $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $t \ge 0$,

$$\mathcal{A}_{r,j}^N(\theta_j^N, t) = \int_0^t \int_0^{F_{j,t}^N(\tilde{\chi}_j^N(u))} h_j^r\left(\left(F_{j,u}^N\right)^{-1}(y)\right) dy du, \tag{62}$$

$$\mathcal{A}_{r,j}^N(\Theta_j^N, t) = \int_0^t \int_0^{F_{j,t}^N(\chi_j^N(u-))} h_j^r\left(\left(F_{j,u}^N\right)^{-1}(y)\right) dy du. \tag{63}$$

Proof. Fix $N \in \mathbb{N}$ and $j \in \mathbb{J}$. It suffices to verify (61) because, for each $t \ge 0$, (62) and (63) follow by respectively taking $x = x(u) = \tilde{\chi}_j^N(u)$ and $x = x(u) = \chi_j^N(u-)$ for each $u \in [0, t]$ in (61), integrating over $[0, t]$, and using (60) to obtain (62). Hence, we proceed with the verification of (61). Fix $t \ge 0$. If $\langle 1, \eta_j^N(t)\rangle = 0$, (61) holds trivially. Henceforth, we assume that $\langle 1, \eta_j^N(t)\rangle > 0$. Given $1 \le k \le \langle 1, \eta_j^N(t)\rangle$, let $\tilde{w}_{j,k}(t) = \inf\{x \in \mathbb{R}_+ : F_{j,t}^N(x) \ge k\}$, which is the k^{th} largest waiting time among class j customers that have entered the system but not reneged by time t. Then $0 \le \tilde{w}_{j,1}(t) \le \tilde{w}_{j,2}(t) \le \cdots \le \tilde{w}_{j,\langle 1,\eta_j^N(t)\rangle}(t)$. The inequalities are weak because E^N may have jumps of finite but arbitrary size. By definition of $F_{j,t}^N$, for $y \in \left[0, \langle 1, \eta_j^N(t)\rangle\right]$,

$$\left(F_{j,t}^N\right)^{-1}(y) = \inf\left\{x \in \mathbb{R}_+ : F_{j,t}^N(x) \ge y\right\} = \inf\left\{x \in \mathbb{R}_+ : F_{j,t}^N(x) \ge \lceil y\rceil\right\} = \left(F_{j,t}^N\right)^{-1}(\lceil y\rceil) = \tilde{w}_{j,\lceil y\rceil}^N(t).$$

Therefore, for $k = 1, \ldots, \langle 1, \eta_j^N(t)\rangle$,

$$h_j^r(\tilde{w}_{j,k}^N(t)) = h_j^r\left(\left(F_{j,t}^N\right)^{-1}(k)\right) = h_j^r\left(\left(F_{j,t}^N\right)^{-1}(k)\right)\int_{k-1}^k dy = \int_{k-1}^k h_j^r\left(\left(F_{j,t}^N\right)^{-1}(y)\right) dy.$$

Then, (61) follows because, for each $x \in \mathbb{R}_+$, $\langle 1_{[0,x]} h_j^r, \eta_j^N(t)\rangle = \sum_{k=1}^{F_{j,t}^N(x)} h_j^r(\tilde{w}_{j,k}^N(t))$. □

4.1.2. Finite Radon Measure Interpretation. Observe that for each $j \in \mathbb{J}$ and $t \geq 0$, $\mathcal{D}_j^N(\cdot,t)$ and $\mathcal{A}_{s,j}^N(\cdot,t)$ (resp. $\mathcal{S}_j^N(\cdot,t)$ and $\mathcal{A}_{r,j}^N(\cdot,t)$) are nonnegative linear functionals on $\mathbf{C}_c([0,H_j^s)\times\mathbb{R}_+)$ (resp. $\mathbf{C}_c([0,H_j^r)\times\mathbb{R}_+)$). This together with (21) and (22) implies that for each $j \in \mathbb{J}$ and $t \geq 0$, $\mathcal{D}_j^N(\cdot,t)$ and $\mathcal{S}_j^N(\cdot,t)$ are finite Radon measures on $[0,H_j^s)\times\mathbb{R}_+$ and $[0,H_j^r)\times\mathbb{R}_+$, respectively. Hence, for each $j \in \mathbb{J}$ and $t \geq 0$, $\mathcal{D}_j^N(\cdot,t) \in \mathbf{M}([0,H_j^s)\times\mathbb{R}_+)$ and $\mathcal{S}_j^N(\cdot,t) \in \mathbf{M}([0,H_j^r)\times\mathbb{R}_+)$. As a result of the next lemma, we have the analogous properties for the compensator processes.

Lemma 7. *For each $N \in \mathbb{N}$, $j \in \mathbb{J}$, $0 < m < H_j^s$, $t \geq 0$, and bounded measurable $\varphi : [0,H_j^s)\times\mathbb{R}_+ \to \mathbb{R}$ such that* $\operatorname{supp}(\varphi) \subseteq [0,m]\times\mathbb{R}_+$,

$$\left|\mathcal{A}_{s,j}^N(\varphi,t)\right| \leq \|\varphi\|_\infty\left(X_j^N(0)+E_j^N(t)\right)\int_0^m h_j^s(x)dx.$$

Similarly, for each $N \in \mathbb{N}$, $j \in \mathbb{J}$, $0 < m < H_j^r$, $t \geq 0$, and bounded measurable $\psi : [0,H_j^r)\times\mathbb{R}_+ \to \mathbb{R}$ such that $\operatorname{supp}(\psi) \subseteq [0,m]\times\mathbb{R}_+$,

$$\left|\mathcal{A}_{r,j}^N(\psi,t)\right| \leq \|\psi\|_\infty\left(\langle 1,\eta_j^N(0)\rangle+E_j^N(t)\right)\int_0^m h_j^r(x)dx.$$

Lemma 7 implies that for each $j \in \mathbb{J}$ and $t \geq 0$, $\mathcal{A}_{s,j}^N(\cdot,t)$ and $\mathcal{A}_{r,j}^N(\cdot,t)$ are finite, nonnegative Radon measures. In particular, for each $j \in \mathbb{J}$ and $t \geq 0$, $\mathcal{A}_{s,j}^N(\cdot,t) \in \mathbf{M}([0,H_j^s)\times\mathbb{R}_+)$ and $\mathcal{A}_{r,j}^N(\cdot,t) \in \mathbf{M}([0,H_j^r)\times\mathbb{R}_+)$. In addition, for each $j \in \mathbb{J}$ and $t \geq 0$, $\mathcal{M}_{s,j}^N(\cdot,t)$ and $\mathcal{M}_{r,j}^N(\cdot,t)$ are finite, signed Radon measures. We will find it convenient to write $\mathcal{D}^N$, $\mathcal{A}_s^N$, $\mathcal{M}_s^N$, $\mathcal{S}^N$, $\mathcal{A}_r^N$, and $\mathcal{M}_r^N$ to denote the vector measure valued processes with respective jth coordinates given by $\mathcal{D}_j^N$, $\mathcal{A}_{s,j}^N$, $\mathcal{M}_{s,j}^N$, $\mathcal{S}_j^N$, $\mathcal{A}_{r,j}^N$, and $\mathcal{M}_{r,j}^N$ for $j \in \mathbb{J}$.

4.2. Tightness

Let

$$\mathcal{V} = (\mathbf{D}(\mathbb{R}_+^J))^2 \times \times_{j=1}^J \mathbf{D}(\mathbf{M}[0,H_j^s)) \times \times_{j=1}^J \mathbf{D}(\mathbf{M}[0,H_j^r)) \times (\mathbf{D}(\mathbb{R}_+^J))^6 \times (\mathbf{D}(\mathbb{R}_+))^2$$
$$\times (\times_{j=1}^J \mathbf{D}(\mathbf{M}([0,H_j^s)\times\mathbb{R}_+))^2 \times (\times_{j=1}^J \mathbf{D}(\mathbf{M}([0,H_j^r)\times\mathbb{R}_+))^2,$$

which is endowed with the product metric. Then $\mathcal{V}$ is a product of Polish spaces and is therefore a Polish space.

Theorem 2 (Tightness). *Suppose that Assumptions 1, 2, and 3 hold. For each $N \in \mathbb{N}$, let*

$$\bar{V}^N = \left(\bar{E}^N,\ \bar{X}^N,\ \bar{\nu}^N,\ \bar{\eta}^N,\ \langle 1,\bar{\eta}^N\rangle,\ \bar{B}^N,\ \bar{Q}^N,\ \bar{R}^N,\bar{D}^N,\ \bar{K}^N,\ \bar{I}^N,\ \bar{D}_\Sigma^N,\ \bar{\mathcal{D}}^N,\ \bar{\mathcal{A}}_s^N,\ \bar{\mathcal{S}}^N,\ \bar{\mathcal{A}}_r^N\right).$$

Then $\{\bar{V}^N\}_{N\in\mathbb{N}}$ is relatively compact in $\mathcal{V}$ and is therefore tight.

Theorem 2 follows from Lemmas 8 and 9. The proofs of Lemmas 8 and 9 leverage many aspects of the proofs of relative compactness as developed in Kang and Ramanan [11] and Kaspi and Ramanan [12]. The proof of relative compactness given in Kaspi and Ramanan [12, section 5] is for the single-class case without reneging. Reneging is incorporated into the single-class model in Kang and Ramanan [11]. The aspects of these statements and proofs that extend in a relatively straightforward fashion are summarized in the appendix. Here we focus on the main contribution of the present paper, which is to handle the multiclass case with control. We note that multiclass many-server queues operating under static priority disciplines are treated in Atar et al. [4] also using elements of the proofs in Kang and Ramanan [11] and Kaspi and Ramanan [12] together with a Skorokhod mapping developed in Atar et al. [4] that is specific to static priority policies. To begin their argument, Atar et al. [4] note that the relative compactness of $\{\bar{D}^N\}_{N\in\mathbb{N}}$ and $\{\bar{R}^N\}_{N\in\mathbb{N}}$ follow precisely as the case treated in Kang and Ramanan [11, lemma 6.3]. Because these arguments do not use the specifics of the control policy, this is also true in our case, as described more fully in the appendix. However, because we want to prove tightness for any control policy, our proof departs from the line of reasoning in Atar et al. [4] thereafter. As previously mentioned, the line of reasoning used here follows more closely to that used in Kang and Ramanan [11] and Kaspi and Ramanan [12]; but some adaptations need to be made to account for the fact that our control policies are multiclass and may idle at times when there is work in certain buffers. This is accounted for in the proof of the next lemma, Lemma 8, which is a multiclass analog of Kang and Ramanan [11, lemmas 6.3 and 6.4].

Lemma 8. *Suppose that Assumptions 1, 2, and 3 hold. For each* $j \in \mathbb{J}$, $f_j \in \mathbf{C}_c^1([0,H_j^s))$, $\varsigma_j \in \mathbf{C}_c^1([0,H_j^r))$, $\varphi_j \in \mathbf{C}_b([0,H_j^s)\times\mathbb{R}_+)$, *and* $\psi_j \in \mathbf{C}_b([0,H_j^r)\times\mathbb{R}_+)$, *the sequences* $\{\bar{E}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{X}_j^N\}_{N\in\mathbb{N}}$, $\{\langle f_j,\bar{\nu}_j^N\rangle\}_{N\in\mathbb{N}}$, $\{\langle \varsigma_j,\bar{\eta}_j^N\rangle\}_{N\in\mathbb{N}}$, $\{\langle 1,\bar{\eta}_j^N\rangle\}_{N\in\mathbb{N}}$, $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{Q}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{R}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{D}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{D}}_j^N(\varphi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{A}}_{s,j}^N(\varphi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{S}}_j^N(\psi_j,\cdot)\}_{N\in\mathbb{N}}$, *and* $\{\bar{\mathcal{A}}_{r,j}^N(\psi_j,\cdot)\}_{N\in\mathbb{N}}$ *are relatively compact in* $\mathbf{D}(\mathbb{R}_+)$ *and are therefore tight. Similarly, the sequences* $\{\bar{I}^N\}_{N\in\mathbb{N}}$ *and* $\{\bar{D}_\Sigma^N\}_{N\in\mathbb{N}}$ *are relatively compact in* $\mathbf{D}(\mathbb{R}_+)$ *and are therefore tight. If Assumption 5 also holds, then each of these processes is* **C***-tight.*

The proof of Lemma 8 uses Kurtz' criteria, which are stated in Section A.1 in the appendix.

Proof of Lemma 8. For $j \in \mathbb{J}$, fix $\varphi_j \in \mathbf{C}_b([0,H_j^s)\times\mathbb{R}_+)$, $\psi_j \in \mathbf{C}_b([0,H_j^r)\times\mathbb{R}_+)$, $f_j \in \mathbf{C}_c^1([0,H_j^s))$, and $\varsigma_j \in \mathbf{C}_c^1([0,H_j^r))$. Because of Part 1 of Assumption 1, $\{\bar{E}_j^N\}_{N\in\mathbb{N}}$ is relatively compact for each $j \in \mathbb{J}$. Relative compactness of the sequences $\{\bar{\mathcal{D}}_j^N(\varphi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{A}}_{s,j}^N(\varphi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{S}}_j^N(\psi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{A}}_{r,j}^N(\psi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{D}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{R}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{X}_j^N\}_{N\in\mathbb{N}}$, and $\{\langle 1,\bar{\eta}_j^N\rangle\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$ follow exactly as in Kang and Ramanan [11, lemma 6.3] by adding class subscripts and applying Lemma 4 and (57)–(58) in place of Kang and Ramanan [11, proposition 5.1(1)] and Lemma A.1(2) in place of Kaspi and Ramanan [12, lemma 5.8(2)] and Kang and Ramanan [11, remark 5.2]. Therefore, we do not reproduce those arguments here. Instead, we simply make note that Assumption 2 is not needed for these arguments. It follows that $\{\bar{D}_\Sigma^N\}_{N\in\mathbb{N}}$ is also relatively compact. Because each of these processes take values in a Polish space, relative compactness is equivalent to tightness; so each of these processes is tight. In addition, the sequences $\{\bar{\mathcal{D}}_j^N(\varphi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{A}}_{s,j}^N(\varphi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{A}}_{r,j}^N(\psi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{D}_j^N\}_{N\in\mathbb{N}}$, and $\{\bar{R}_j^N\}_{N\in\mathbb{N}}$ are **C**-tight because each process in these sequences is either continuous or has jumps of size $1/N$. Because we allowed for batch arrivals, the processes in the sequences $\{\bar{E}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{S}}_j^N(\psi_j,\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{X}_j^N\}_{N\in\mathbb{N}}$, and $\{\langle 1,\bar{\eta}_j^N\rangle\}_{N\in\mathbb{N}}$ may have jumps larger than size $1/N$. However, when parts 1 and 2 of Assumption 5, it is straightforward to verify that each of these sequences is **C**-tight.

Next we argue that the remaining $\mathbf{D}(\mathbb{R}_+)$ valued processes $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{Q}_j^N\}_{N\in\mathbb{N}}$, and $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$, for each $j \in \mathbb{J}$, and $\{\bar{I}^N\}_{N\in\mathbb{N}}$ are relatively compact. First we show that they satisfy (K.1) of Kurtz' criteria. Clearly, $\{\bar{I}^N\}_{N\in\mathbb{N}}$ satisfies (K.1). By (15), nonnegativity of $\bar{B}^N$ and $\bar{Q}^N$ and (19), for each $j \in \mathbb{J}$ and $t \geq 0$,

$$0 \leq \bar{B}_j^N(t) \vee \bar{Q}_j^N(t) \leq \bar{X}_j^N(t) \qquad \text{and} \qquad \bar{K}_j^N(t) \leq \bar{B}_j^N(t) + \bar{D}_j^N(t).$$

For each $j \in \mathbb{J}$, $\{\bar{X}_j^N\}_{N\in\mathbb{N}}$ is relatively compact and so it satisfies (K.1). Therefore, $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$ and $\{\bar{Q}_j^N\}_{N\in\mathbb{N}}$ satisfy (K.1) for each $j \in \mathbb{J}$. For each $j \in \mathbb{J}$, $\{\bar{D}_j^N\}_{N\in\mathbb{N}}$ is relatively compact and so it satisfies (K.1). Therefore, $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$ satisfies (K.1) for each $j \in \mathbb{J}$.

We continue by arguing that each of the $\mathbf{D}(\mathbb{R}_+)$ valued processes $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{Q}_j^N\}_{N\in\mathbb{N}}$, and $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$, for each $j \in \mathbb{J}$, and $\{\bar{I}^N\}_{N\in\mathbb{N}}$ satisfy (K.2) of Kurtz' criteria and are therefore relatively compact. By Assumption 2 and the relative compactness of $\{\bar{E}_j^N\}_{N\in\mathbb{N}}$ and $\{\bar{D}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$, $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$ satisfies (K.2) for each $j \in \mathbb{J}$. Then, by (19) and the relative compactness of $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$ and $\{\bar{D}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$, $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$ satisfies (K.2) for each $j \in \mathbb{J}$. In turn, (16) and the relative compactness of $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$ imply that $\{\bar{I}^N\}_{N\in\mathbb{N}}$ satisfies (K.2). Similarly, (15) and the relative compactness of $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$ and $\{\bar{X}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$ imply that $\{\bar{Q}_j^N\}_{N\in\mathbb{N}}$ satisfies (K.2) for each $j \in \mathbb{J}$. Therefore, each of these $\mathbf{D}(\mathbb{R}_+)$ valued processes is relatively compact and therefore tight. If Assumption 5 holds, then, as noted in the first paragraph of this proof, Assumptions 5.1 and 5.2 imply that $\{\bar{E}_j^N\}_{N\in\mathbb{N}}$ and $\{\bar{D}_j^N\}_{N\in\mathbb{N}}$ are **C**-tight for each $j \in \mathbb{J}$, and Assumption 5.3 and **C**-tightness of $\{\bar{E}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$ and of $\{\bar{D}_\Sigma^N\}_{N\in\mathbb{N}}$ imply **C**-tightness of $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$. This together with (15), (16), and (19) in turn imply **C**-tightness of $\{\bar{B}_j^N\}_{N\in\mathbb{N}}$ and $\{\bar{Q}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$ and of $\{\bar{I}^N\}_{N\in\mathbb{N}}$.

That the sequences $\{\langle f_j,\bar{\nu}_j^N\rangle\}_{N\in\mathbb{N}}$ and $\{\langle \varsigma_j,\bar{\eta}_j^N\rangle\}_{N\in\mathbb{N}}$ are tight for each $j \in \mathbb{J}$ follows similarly to the arguments given in the proof of Kang and Ramanan [11, lemma 6.4] by adding subscripts to account for class and using (24) and (26) in Lemma 1 in place of Kang and Ramanan [11, equations (2.27) and (2.28)] or Kaspi and Ramanan [12, equation (5.4)]. Thus, the details are omitted. Finally, suppose that Assumption 5 holds. Then, because $f_j' \in \mathbf{C}_c([0,H_j^s))$ and $\varsigma_j' \in \mathbf{C}_c([0,H_j^r))$ for each $j \in \mathbb{J}$, $\{\mathcal{D}_j(f_j',\cdot)\}_{N\in\mathbb{N}}$ and $\{\mathcal{S}_j(\varsigma_j',\cdot)\}_{N\in\mathbb{N}}$ are **C**-tight for each $j \in \mathbb{J}$, as noted in the first paragraph of this proof. Then **C**-tightness of $\{\langle f_j,\bar{\nu}_j^N\rangle\}_{N\in\mathbb{N}}$ and $\{\langle \varsigma_j,\bar{\eta}_j^N\rangle\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$ follows from (24) and (26) upon utilizing **C**-tightness of $\{\bar{\mathcal{D}}_j^N(f_j',\cdot)\}_{N\in\mathbb{N}}$, $\{\bar{E}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{S}}_j^N(\varsigma_j',\cdot)\}_{N\in\mathbb{N}}$, and $\{\bar{K}_j^N\}_{N\in\mathbb{N}}$ for each $j \in \mathbb{J}$. □

Lemma 9. *Suppose that Assumptions 1, 2, and 3 hold. For each $j \in \mathbb{J}$, the sequences $\{\bar{\nu}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{\eta}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{D}}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{A}}_{s,j}^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{S}}_j^N\}_{N\in\mathbb{N}}$, and $\{\bar{\mathcal{A}}_{r,j}^N\}_{N\in\mathbb{N}}$ are relatively compact in $\mathbf{D}(\mathbf{M}[0,H_j^s))$, $\mathbf{D}(\mathbf{M}[0,H_j^r))$, $\mathbf{D}(\mathbf{M}([0,H_j^s)\times\mathbb{R}_+))$, $\mathbf{D}(\mathbf{M}([0,H_j^s)\times\mathbb{R}_+))$, $\mathbf{D}(\mathbf{M}([0,H_j^r)\times\mathbb{R}_+))$, and $\mathbf{D}(\mathbf{M}([0,H_j^r)\times\mathbb{R}_+))$, respectively.*

Proof. The result follows from Jakubowski's criteria (stated in Section A.1 in the appendix). In particular, by Remark A.1 (stated in Section A.1) and Lemma 8, (J.2) of Jakubowski's criteria holds for each measure valued process. Also, by Lemmas A.3 and A.4 (stated in Section A.3 in the appendix), (J.1) of Jakubowski's criteria holds for each measure valued process. □

4.3. Characterization of Fluid Limit Points

Here we prove Theorem 1. For this, we establish the following lemma.

Lemma 10. *Suppose that Assumptions 4 and 5 hold and that*

$$V = (E, X, \nu, \eta, \langle 1, \eta\rangle, B, Q, R, D, K, I, D_\Sigma, \mathcal{D}, \mathcal{A}_s, \mathcal{S}, \mathcal{A}_r) \tag{64}$$

is a distributional limit point of $\{\bar{V}^N\}_{N\in\mathbb{N}}$, where for each $N \in \mathbb{N}$, $\bar{V}^N$ is as in Theorem 2. Then the following hold almost surely:

1. *for each $j \in \mathbb{J}$, $T > 0$, and $m \in [0, H_j^s)$ (resp. $m \in [0, H_j^s)$), there exists $\tilde{L}_j^s(m,T) < \infty$ (resp. $\tilde{L}_j^r(m,T) < \infty$) such that for each $\ell \in \mathbf{L}^1_{\mathrm{loc}}[0, H_j^s)$ (resp. $\ell \in \mathbf{L}^1_{\mathrm{loc}}[0, H_j^r)$),*

$$\int_0^T \langle \ell, \nu_j(u)\rangle\, du \le \tilde{L}_j^s(m,T) \int_{[0,H_j^s)} |\ell(x)|\, dx \qquad \left(\text{resp.} \int_0^T \langle \ell, \eta_j(u)\rangle du \le \tilde{L}_j^r(m,T) \int_{[0,H_j^r)} |\ell(x)|\, dx\right);$$

2. *for all $j \in \mathbb{J}$, $\varphi \in \mathbf{C}_b([0, H_j^s)\times\mathbb{R}_+)$, $\psi \in \mathbf{C}_b([0, H_j^r)\times\mathbb{R}_+)$, and $t \ge 0$,*

$$\mathcal{D}_j(\varphi, t) = \mathcal{A}_{s,j}(\varphi, t) = \int_0^t \langle \varphi(\cdot, u) h_j^s(\cdot), \nu_j(u)\rangle du < \infty,$$

$$\mathcal{S}_j(\psi, t) = \mathcal{A}_{r,j}(\psi, t) = \int_0^t \langle \psi(\cdot, u) h_j^r(\cdot), \eta_j(u)\rangle du < \infty,$$

and, in particular, (37) and (41) hold;

3. *$(X, \nu, \eta) \in \mathbf{F}$;*
4. *$(E, X, \nu, \eta, R, D, K)$ satisfy (44)–(47);*
5. *the processes X, $\langle 1, \eta\rangle$, B, Q, R, D, K, I, and D_Σ are continuous;*
6. *for all $j \in \mathbb{J}$, $t \ge 0$, and $x \in \mathbb{R}_+$, $\langle 1_{\{x\}}, \eta_j(t)\rangle = 0$;*
7. *R satisfies (40);*
8. *(B, Q, R, D, K, I) are the auxiliary functions for (X, ν, η), that is, they satisfy (38)–(43);*
9. *(X, ν, η) is a fluid model solution for E with $(X(0), \nu(0), \eta(0)) = (X^0, \nu^0, \eta^0)$ that is continuous.*

Theorem 1 immediately follows from Lemma 10 as follows.

Proof of Theorem 1. Let (X, ν, η) be a distributional limit point of $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N\in\mathbb{N}}$. Then there exists $\{N'\}$, a subsequence of $\{N\}$, such that $(\bar{X}^{N'}, \bar{\nu}^{N'}, \bar{\eta}^{N'}) \Rightarrow (X, \nu, \eta)$ as $N' \to \infty$. Let $\{\bar{V}^N\}_{N\in\mathbb{N}}$ be as defined in Theorem 2 and consider the subsequence $\{\bar{V}^{N'}\}$. Let $\tilde{V}$ be a limit point of $\{\bar{V}^{N'}\}$. Then there exists $\{N''\}$, a subsequence of $\{N'\}$, such that $\bar{V}^{N''} \Rightarrow \tilde{V}$ as $N \to \infty$. So then, because $\bar{X}^{N''}$, $\bar{\nu}^{N''}$, and $\bar{\eta}^{N''}$ are coordinates of $\bar{V}^{N''}$, it follows that $(\tilde{X}, \tilde{\nu}, \tilde{\eta})$ is equal in distribution to (X, ν, η). Furthermore, by Lemma 10, $(\tilde{X}, \tilde{\nu}, \tilde{\eta})$ is almost surely a fluid model solution that is continuous and has continuous auxiliary functions such that $\tilde{\eta}_j(t)$ does not charge points for each $t \ge 0$ and $j \in \mathbb{J}$. Hence, the same is true for (X, ν, η). □

Proof of Lemma 10. Let $\{\bar{V}^N\}_{N\in\mathbb{N}}$ be as defined in Theorem 2. Suppose that $\{N'\}$ is a subsequence of $\{N\}$ such that, as $N' \to \infty$, $\bar{V}^{N'} \Rightarrow V$, where V is given by (64). To ease the notation, we denote this subsequence by $\{\bar{V}^N\}$ and invoke the Skorokhod representation theorem so that we may assume that $\bar{V}^N \to V$ almost surely as $N \to \infty$.

We begin by verifying Lemma 10(1). For each $j \in \mathbb{J}$, the proof for $\bar{\nu}_j$ (resp. $\bar{\eta}_j$) follows analogously to the proof of Kaspi and Ramanan [12, lemma 5.16] (resp. Kang and Ramanan [11, lemma 7.4]).

Next we verify Lemma 10(2). Lemma 5 and the converging together lemma imply that for all $j \in \mathbb{J}$, $\varphi \in \mathbf{C}_b([0,H_j^s)\times\mathbb{R}_+)$, $\psi \in \mathbf{C}_b([0,H_j^r)\times\mathbb{R}_+)$, $\mathcal{D}_j(\varphi,\cdot)=\mathcal{A}_{s,j}(\varphi,\cdot)$, and $\mathcal{S}_j(\psi,\cdot)=\mathcal{A}_{s,j}(\psi,\cdot)$ for each $j \in \mathbb{J}$. Therefore, it suffices to establish the second equality in each expression. The verification of this can be done by using essentially the same argument given to verify Kaspi and Ramanan [12, (5.49) in proposition 5.17], except with the inclusion of subscripts to account for class and using Lemma A.1(1) in place of Kaspi and Ramanan [12, lemma 5.8(1)], Proposition A.1 in place of Kaspi and Ramanan [12, proposition 5.7], and the already established part (1) of this lemma (Lemma 10(1)) in place of Kaspi and Ramanan [12, lemma 5.16]. Therefore, the details are omitted.

To demonstrate that Lemma 10(3) holds, we must show that (35), (36), and (37) hold almost surely. The fact that (35) and (36) hold follows from $\bar{V}^N \to V$ almost surely as $N \to \infty$, (S.2), (13), (15), and (28). The already established part (2) of this lemma (Lemma 10(2)) shows that (37) holds.

Next we verify Lemma 10(4). The requirement in (44) that K_j is nondecreasing for each $j \in \mathbb{J}$ follows from $\bar{V}^N \to V$ almost surely as $N \to \infty$ and that K_j^N is a counting process for each $j \in \mathbb{J}$ and $N \in \mathbb{N}$ (and, therefore, nondecreasing). The balance Equation (45) holds almost surely from $\bar{V}^N \to V$ and the system balance Equation (17). Arguments very similar to those used to establish Kaspi and Ramanan [12, (3.5) for theorem 5.15 (see page 104)] and Kang and Ramanan [11, (3.11) for theorem 7.1 (see page 51)] also respectively show that Equations (48) and (49) hold almost surely, which are equivalent to (46) and (47) because (37) and (44) hold (see Remark 1).

Lemma 10(5) follows from the fact that the prelimit processes are **C**-tight as shown in Lemma 8.

Next we verify Lemma 10(6). Given $j \in \mathbb{J}$, $t \geq 0$, and $x \in \mathbb{R}_+$, because $1_{\{x\}}$ is a bounded Borel measurable function, (47) implies that

$$\langle 1_{\{x\}}, \eta_j(t)\rangle = \begin{cases} \dfrac{1-G_j^r(x)}{1-G_j^r(x-t)}\langle 1_{\{x-t\}}, \eta_j^0\rangle, & \text{if } x \geq t, \\ \Big(1-G_j^r(t-x)\Big)\Big(E_j(t-x)-E_j((t-x)-)\Big), & \text{if } x < t. \end{cases}$$

By Assumption 5, the right side is zero in both cases.

Next we verify Lemma 10(7). To facilitate this, for $j \in \mathbb{J}$ and $t \geq 0$, let

$$\tilde{R}_j(t) = \int_0^t \int_0^{Q_j(u)} h_j^r\left(\left(F_{j,u}\right)^{-1}(y)\right) dy du. \tag{65}$$

For each $N \in \mathbb{N}$, $j \in \mathbb{J}$, $u \geq 0$, and $x \in [0,H_j^r)$, let $\bar{F}_{j,u}^N(x) = F_{j,u}^N(x)/N$. Then, for each $N \in \mathbb{N}$, $j \in \mathbb{J}$, $u \geq 0$, and $x \in [0,H_j^r)$, $\left(\bar{F}_{j,u}^N\right)^{-1}(x) = \left(F_{j,u}^N\right)^{-1}(Nx)$. This together with (62) and (63) in Lemma 6, implies that for each $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $t \geq 0$,

$$\bar{\mathcal{A}}_{r,j}^N(\Theta_j^N, t) = \int_0^t \int_0^{\bar{F}_{j,u}^N(\chi_j^N(u-))} h_j^r\left(\left(\bar{F}_{j,u}^N\right)^{-1}(y)\right) dy du, \tag{66}$$

$$\bar{\mathcal{A}}_{r,j}^N(\theta_j^N, t) = \int_0^t \int_0^{\bar{F}_{j,u}^N(\chi_j^N(u))} h_j^r\left(\left(\bar{F}_{j,u}^N\right)^{-1}(y)\right) dy du. \tag{67}$$

Similarly to Kang and Ramanan [11, the arguments in the proof of proposition 7.2], we will verify that, for all $T \geq 0$,

$$\lim_{N\to\infty} \mathbb{E}\left[\max_{j\in\mathbb{J}} \sup_{t\in[0,T]} \left|\bar{\mathcal{A}}_{r,j}^N(\Theta_j^N, t) - \tilde{R}_j(t)\right| \vee \max_{j\in\mathbb{J}} \sup_{t\in[0,T]} \left|\bar{\mathcal{A}}_{r,j}^N(\theta_j^N, t) - \tilde{R}_j(t)\right|\right] = 0. \tag{68}$$

Then (68) together with Lemma 5 and the converging together lemma implies that for each $t \geq 0$ and $j \in \mathbb{J}$, $\bar{\mathcal{S}}_{r,j}^N(\Theta_j^N, t)$ and $\bar{\mathcal{S}}_{r,j}^N(\theta_j^N, t)$ converge in probability to $\tilde{R}_j(t)$ as $N \to \infty$. Also, for each $j \in \mathbb{J}$ and $t \geq 0$, $\bar{R}_j^N(t)$ converges to $R_j(t)$ almost surely as $N \to \infty$. Hence, for each $j \in \mathbb{J}$ and $t \geq 0$, $R_j(t) = \tilde{R}_j(t)$ almost surely. Because R_j is almost surely continuous (because of **C**-tightness of $\{\bar{R}_j^N\}_{N=1}^\infty$) and $\tilde{R}_j$ is also almost surely continuous (because of (65)), $R_j = \tilde{R}_j$ almost surely and (40) holds.

It remains to justify (68). To begin we show that, almost surely, for all $j \in \mathbb{J}$ and $t \geq 0$,

$$\lim_{N\to\infty} \bar{F}_{j,t}^N(\tilde{\chi}_j^N(t)) = \lim_{N\to\infty} \bar{F}_{j,t}^N(\chi_j^N(t-)) = Q_j(t). \tag{69}$$

For this, for $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $t \geq 0$, observe that

$$\bar{Q}_j^N(t) - \left\langle 1_{\{\chi_j^N(t-)\}}, \bar{\eta}_j^N(t)\right\rangle \leq \bar{F}_{j,t}^N(\tilde{\chi}_j^N(t)) \leq \bar{F}_{j,t}^N(\chi_j^N(t-)) \leq \bar{Q}_j^N(t) + \left\langle 1_{\{\chi_j^N(t-)\}}, \bar{\eta}_j^N(t)\right\rangle.$$

Furthermore, note that for $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $t \geq 0$,

$$\left\langle 1_{\{\chi_j^N(t-)\}}, \bar{\eta}_j^N(t) \right\rangle \leq (\bar{E}_j^N((t-\chi_j^N(t-))^+) - \bar{E}_j^N((t-\chi_j^N(t-))^+-)) \vee \left\langle 1_{\{(\chi_j^N(t-)-t)^+\}}, \bar{\eta}_j^N(0) \right\rangle,$$
$$\leq \sup_{0\leq u\leq t} (\bar{E}_j^N(u) - \bar{E}_j^N(u-)) \vee \sup_{x\in\mathbb{R}_+} \langle 1_{\{x\}}, \bar{\eta}_j^N(0)\rangle.$$

Upon letting $N \to \infty$ and using Assumption 5 and the already established part (6) of this lemma (Lemma 10), it follows that almost surely, for each $j \in \mathbb{J}$ and $t \geq 0$, the right-hand side of the previous display converges to zero. Thus, (69) holds. Given that (69) holds, (68) follows by arguing similarly to the proof of Kang and Ramanan [11, proposition 7.2]. To verify this, for each $j \in \mathbb{J}$ and $u \geq 0$, one should add class-level subscripts, use (66) and (67) in place of Kang and Ramanan [11, (7.5)], replace the limit of integration $\bar{Q}^{(N)}(u) + \bar{i}^{(N)}(u)$ throughout the remainder of that argument with $\bar{F}_{j,u}^N(\chi_j^N(u-))$ or $\bar{F}_{j,u}^N(\tilde{\chi}_j^N(u))$ as appropriate, use the already established parts (1) and (6) of this lemma (Lemma 10) in place of Kang and Ramanan [11, lemma 7.4] and Kang and Ramanan [11, lemma 7.3] respectively, and use (69) in place of showing $\bar{Q}^{(N)}(u) + \bar{i}^{(N)}(u)$ converges to $Q(u)$ for almost every $u \geq 0$. We remark that the fact that E is continuous and the coordinates of η_j^0 do not charge points for each $j \in \mathbb{J}$ makes it possible to simplify some of the arguments.

Next we verify Lemma 10(8). The already established parts (3) and (7) of this lemma (Lemma 10) imply that (41) and (40), respectively, hold almost surely. Equations (38), (39), (42), and (43) hold almost surely by (13), (16), (15), and (19) and the fact that V is the almost sure limit of $\{\bar{V}^N\}_{N\in\mathbb{N}}$.

Finally, Lemma 10(9) follows from the already established parts (3), (4), (5), and (8) of this lemma (Lemma 10) because Assumption 3 guarantees that $(X(0), \nu(0), \eta(0)) = (X^0, \nu^0, \eta^0)$ almost surely, and from Lemma 2(i). □

4.4. Application of Theorem 1

As previously mentioned, if one has a specific collection of admissible HL control policies that they wish to analyze, then Theorem 1 can be used as a step toward proving a fluid limit theorem. In particular, the following additional steps need to be executed:

(P.1) add policy-specific equations to the stochastic model that uniquely characterize the dynamics;

(P.2) add policy-specific equations to the fluid model and, for a given arrival function E and initial condition $(X(0), \nu(0), \eta(0))$, prove uniqueness of fluid model solutions that also satisfy the policy-specific fluid model equations;

(P.3) prove that fluid limit points of sequences satisfying the assumptions of Theorem 1 and the stochastic policy-specific equations also satisfy the policy-specific fluid model equations.

Together (P.1)–(P.3) and Theorem 1 imply uniqueness in law of the limits points and thus convergence.

To illustrate this for an established example, consider the static priority policy in Atar et al. [4]. The policy-specific equations for the N-server queue with static priority control are given by Atar et al. [4, equations (8) and (16)]: for $t \geq 0$,

$$N - \sum_{j=1}^J B_j^N(t) = \left(N - \sum_{j=1}^J X_j^N(t)\right)^+, \tag{70}$$

$$K_j^N(t) = \int_{[0,t]} 1_{\left\{\sum_{i=1}^{j-1} Q_i^N(s)=0\right\}} dK_j^N(s), \qquad 2 \leq j \leq J. \tag{71}$$

Observe that (5)–(26), (70), and (71) uniquely determine an HL control policy for the N-server queue as in Definition 1, which we refer to as static priority. Furthermore, for a suitable initial condition, the static priority policy is an admissible HL control policy as in Definition 3. Thus, (P.1) is accomplished for static priority.

The additional policy-specific equations that static priority fluid model solutions must satisfy are given by Atar et al. [4, equations (30) and (32)]: for $t \geq 0$,

$$1 - \sum_{j=1}^J B_j(t) = \left(1 - \sum_{j=1}^J X_j(t)\right)^+, \tag{72}$$

$$K_j(t) = \int_{[0,t]} 1_{\left\{\sum_{i=1}^{j-1} Q_i(s)=0\right\}} dK_j(s), \qquad 2 \leq j \leq J. \tag{73}$$

We refer to fluid model solutions for an arrival function E that also satisfy (72) and (73) as static priority fluid model solutions for E. Under the assumption that the hazard rates for each of the reneging distributions are bounded, static priority fluid model solutions are shown to be unique in Atar et al. [4] (see Atar et al. [4, theorem 3.1]). Thus, (P.2) is accomplished for static priority.

It remains to verify (P.3). Assume that the hazard rates for each of the reneging distributions are bounded. Let $\{Y^N\}_{N\in\mathbb{N}}$ be a state sequence for an admissible static priority HL control. Then (55) holds. Suppose that Assumption 1 holds for some E that is continuous and that Assumption 3 holds for some η^0 such that η^0_j does not charge points for all $j \in \mathbb{J}$. Then Assumption 5 holds. Further suppose that Assumption 4 holds. Hence, by Theorem 1, $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N\in\mathbb{N}}$ is tight and any limit point is a fluid model solution almost surely. Fix a limit point (X, ν, η). In order to complete (P.3), we must show that (X, ν, η) also satisfies (72) and (73). That (X, ν, η) satisfies (72) follows by applying the Skorokhod representation theorem to assume that the subsequence converges to (X, ν, η) almost surely and using the fact that Y^N satisfies (70) for each $N \in \mathbb{N}$. So it suffices to verify (73). This is done within the proof of Atar et al. [4, theorem 4.3] in the slightly more restrictive setting where the prelimit arrival processes are renewal processes with interarrival distributions that have densities and such that the limiting arrival function E satisfies $E_j(t) = \lambda_j t$ for each class $j \in \mathbb{J}$ and some $\lambda \in (0,\infty)^J$. They also assume that G^r_L is strictly increasing, where $L := \inf\{j : \sum_{i=1}^{j} \lambda_i m_i \geq 1\}$ is the index of the class that could be partially served, if such an L exists. The approach there using an appropriately chosen Skorokhod mapping problem can be carried out under the less restrictive assumption that E is continuous, as assumed here. So (X, ν, η) satisfies the policy-specific Equations (72) and (73) and (P.3) holds. This together with the fact that $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N\in\mathbb{N}}$ is tight and all fluid limit points of $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N\in\mathbb{N}}$ are fluid model solutions implies that all fluid limit points of $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N\in\mathbb{N}}$ are static priority fluid model solutions for the arrival function E. Hence, assuming that h^r_j is bounded for each $j \in \mathbb{J}$, convergence follows from (P.2).

We anticipate that such a program could be followed for many admissible HL policies of interest. In Section 5, we give a second illustration for an admissible HL control policy not previously studied.

5. Weighted Random Buffer Selection

In this section, we follow the program outlined in Section 4.4 to study a subset of admissible HL control policies that we refer to as weighted random buffer selection policies. We begin in Section 5.1 by formally defining WRBS policies. The policy-specific equations for the stochastic model are given in (74) and (75). In Section 5.2, WRBS policy-specific fluid model Equations (77) and (78) are added to the fluid model equations to specify WRBS fluid model solutions. The main result for WRBS fluid model solutions is the uniqueness result stated as Theorem 3 in Section 5.2.2. We also show continuity properties of WRBS fluid model solutions in Lemma 11 in Section 5.2.1 and characterize the invariant states of the WRBS policy class in Theorem 4 in Section 5.2.3. In Section 5.3, we state and prove the final result of this section, Theorem 5, which is a fluid limit theorem for WRBS policies.

5.1. Weighted Random Buffer Selection Policies

Fix weights $p_j \geq 0$ with $\sum_{j=1}^{J} p_j = 1$. For $i \in \mathbb{N}$, let j^*_i be the unique $j \in \mathbb{J}$ such that

$$\sum_{k=1}^{j-1} p_k < d^N_i \leq \sum_{k=1}^{j} p_k.$$

Then $\{j^*_i\}_{i\in\mathbb{N}}$ is an i.i.d. sequence of random variables such that $\mathbb{P}(j^*_1 = j) = p_j$ for each $j \in \mathbb{J}$. Whenever a departure occurs, the next random variable in the list $\{j^*_i\}_{i\in\mathbb{N}}$ is used to randomly select a customer class or buffer. Specifically, if $u \geq 0$ is such that $D^N_\Sigma(u) - D^N_\Sigma(u-) = 1 > 0$, then there is a departure at time u and the value of the random variable $j^*_{D^N_\Sigma(u)}$ determines which class should send its HL customer into service, if possible. Note that if that class $j^*_{D^N_\Sigma(u)}$ has no customers in queue at time u, it cannot send a customer into service. In this case, a class J nonidling condition is enforced. In particular, if $D^N_\Sigma(u) - D^N_\Sigma(u-) = 1$, $j^*_{D^N_\Sigma(u)} = j < J$ and $Q^N_j(u-) = 0$, the class J HL customer enters service, if possible, that is, if $Q^N_J(u-) > 0$. In addition, class J customers that arrive to find an idle server immediately enter service, which prevents the system from getting stuck in the set of states with no customers in

service. Then, K^N is the vector of counting processes such that for all $t \geq 0$,

$$K_j^N(t) = \int_0^t 1_{\left\{j^*_{D_\Sigma^N(u)}=j, Q_j^N(u-)\geq 1\right\}} dD_\Sigma^N(u), \qquad \text{for } 1 \leq j < J, \tag{74}$$

$$I^N(t) = \left(I^N(t) - Q_J^N(t)\right)^+. \tag{75}$$

Equation (75) enforces that class J is not allowed to hold customers in queue if there is an idle server. In particular, it is equivalent to (32) for $\mathcal{J} = \{J\}$.

The selection of the weights $p_j, j \in \mathbb{J}$ is important to prevent overserving class J, because of (75). In Section 5.2, where we discuss WRBS fluid model solutions, we include a result showing the relationship between the invariant states discussed in Section 3.3 and how to choose the weights $p_j, j \in \mathbb{J}$, so as to achieve a given invariant state (see Section 5.2.3). From that perspective, the fact that classes $1 \leq j < J$ may hold customers in queue when servers are available is for mathematical convenience. Indeed, allowing such idleness facilitates the representation of the queue-length process using the regulator mapping given in Section 3 in Definition 7, which is important for proving the uniqueness of WRBS fluid model solutions in Section 5.2.2 and weak convergence in Section 4.3.

Equations (74) reflect the description of the WRBS dynamic given above for $j < J$. The remainder of the dynamics of WRBS described above result from combining (74) with (75). To see this more explicitly, assume that (74) and (75) hold. Then upon adding (19) over $j \in \mathbb{J}$, solving for $K_J^N(t)$, adding and subtracting N on the right side and replacing the relevant terms with $I^N(0)$ and $I^N(t)$, and using (74), we obtain that for $t \geq 0$,

$$\begin{aligned} K_J^N(t) = &\int_0^t 1_{\{Q_J^N(u-)\geq 1\}} \left(\sum_{j=1}^{J-1} 1_{\left\{j^*_{D_\Sigma^N(u)}=j,\ Q_j^N(u-)=0\right\}} + 1_{\left\{j^*_{D_\Sigma^N(u)}=J\right\}} \right) dD_\Sigma^N(u) \\ &+ \int_0^t 1_{\{Q_J^N(u-)=0\}} \left(\sum_{j=1}^{J-1} 1_{\left\{j^*_{D_\Sigma^N(u)}=j,\ Q_j^N(u-)=0\right\}} + 1_{\left\{j^*_{D_\Sigma^N(u)}=J\right\}} \right) dD_\Sigma^N(u) \\ &+ I^N(0) - I^N(t). \end{aligned}$$

Next express $I^N - I^N(0)$ as the difference of two nondecreasing, nonnegative functions I_+^N and I_-^N so that $I^N(t) - I^N(0) = I_+^N(t) - I_-^N(t)$ for all $t \geq 0$. Note that the number of idle servers may increase only at the time of a departure. This triggers an attempt by WRBS to send a new customer into service. Hence, the second term on the right side of the above is then $I_+^N(t)$. Thus, for $t \geq 0$,

$$K_J^N(t) = \int_0^t 1_{\{Q_J^N(u-)\geq 1\}} \left(\sum_{j=1}^{J-1} 1_{\left\{j^*_{D_\Sigma^N(u)}=j,\ Q_j^N(u-)=0\right\}} + 1_{\left\{j^*_{D_\Sigma^N(u)}=J\right\}} \right) dD_\Sigma^N(u) + I_-^N(t). \tag{76}$$

The first term on the right side of (76) captures class J customers entering service because of a departure event for which WRBS chose class J or a class that had no customers in queue, but class J had at least one customer in queue. Further, observe that the number of idle servers can decrease only at moments of arrivals and ,in the case of WRBS, only at jump times of E_J^N. However, because of (75), I_-^N increases at time u if and only if $E_J^N(u) - E_J^N(u-) > 0$ and $Q_J^N(u-) = 0$. For such a time u, $I_-^N(u) - I_-^N(u-) = \min(N - I^N(u-), E_J^N(u) - E_J^N(u-))$. Thus, the second term corresponds to the dynamic described above whereby class J arrivals enter service upon arrival if there are idle servers.

Let $\mathcal{P}$ denote the set of all $p = (p_1, \ldots, p_J)$ with $p_j \geq 0$ for each $j \in \mathbb{J}$ and $\sum_{i=1}^J p_j = 1$, so that p is a probability measure on $\mathbb{J}$. We refer to members of $\mathcal{P}$ as weighted random buffer selection policies. We refer to $p \in \mathcal{P}$ as WRBS policy p. Observe that the state space for WRBS policies is given by $\mathbb{Y}_{\{J\}}^N$, which we write as $\mathbb{Y}_J^N$ to ease the notation.

5.2. Weighted Random Buffer Selection Fluid Model Solutions

Weighted random buffer selection fluid model solutions are fluid model solutions that also satisfy fluid analogs (74) and (75).

Definition 6. Given an arrival function E and $p \in \mathcal{P}$, a fluid model solution for E and the WRBS policy p is a fluid model solution (X, ν, η) for E such that for each $1 \le j < J$ and $0 \le s < t < \infty$,

$$p_j \int_s^t 1_{\{Q_j(u)>0\}} dD_\Sigma(u) \le K_j(t) - K_j(s) \le p_j \int_s^t dD_\Sigma(u), \tag{77}$$

and

$$I(t) = \big(I(t) - Q_J(t)\big)^+. \tag{78}$$

The two inequalities in (77) are used in lieu of a direct fluid analog of (74) because they are easier to work with in the analysis. A more direct fluid analog of (74) is developed in Section 5.2.1 when the arrival function is absolutely continuous. A uniqueness result is developed in Section 5.2.2 (see Theorem 3).

5.2.1. Continuity Properties of WRBS Fluid Model Solutions.

Lemma 11. *Suppose that $p \in \mathcal{P}$, E is an arrival function, and (X, ν, η) is a fluid model solution for E and WRBS policy p.*

i. *If E is continuous, then (X, ν, η) and each of the auxiliary functions are continuous.*

ii. *If E_j is absolutely continuous with density $\lambda_j(\cdot)$ for each $j \in \mathbb{J}$, then the coordinates of the function X and the coordinates of each of the auxiliary functions (38)–(43) are absolutely continuous with respect to Lebesgue measure. Moreover, for each $1 \le j < J$ and $t \ge 0$,*

$$K_j(t) = \int_0^t \Big(\big(\lambda_j(u) \wedge p_j\delta(u)\big)1_{\{Q_j(u)=0\}} + p_j\delta(u)1_{\{Q_j(u)>0\}}\Big)du, \tag{79}$$

where δ is the density of D_Σ.

Proof. First note that, by (40) and (41), R_j and D_j are absolutely continuous for each $j \in \mathbb{J}$. Then, D_Σ is absolutely continuous. Consequently, by (77), K_j is absolutely continuous for each $1 \le j < J$. And so, by (42), it follows that B_j is absolutely continuous for each $1 \le j < J$.

In order to prove (i), by Lemma 2(i), it suffices to show that K_J is continuous when E is continuous. Similarly, in order to prove the first part of (ii), by Lemma 2(ii), it suffices to show that K_J is absolutely continuous when E_j is absolutely continuous for each $j \in \mathbb{J}$. Suppose that E_j is continuous (resp. absolutely continuous) for each $j \in \mathbb{J}$. Then, by (45), X_j is also continuous (resp. absolutely continuous) for each $j \in \mathbb{J}$. Observe that by rewriting (78) as $I(t) = \Big(1 - \sum_{j=1}^{J-1} B_j(t) - X_J(t)\Big)^+$ for all $t \ge 0$, it follows that I is absolutely continuous (resp. absolutely continuous). Then, by (43), B_J is continuous (resp. absolutely continuous); in turn, by (42), K_J is continuous (resp. absolutely continuous). Thus, (i) and the first part of (ii) hold.

Suppose that E_j is absolutely continuous for each $j \in \mathbb{J}$. To complete the proof, we must verify that (79) holds. Fix $1 \le j < J$ and let κ_j denote the density of K_j. It suffices to show that $\kappa_j(t) = \big(\lambda_j(t) \wedge p_j\delta(t)\big)1_{\{Q_j(t)=0\}} + p_j\delta(t)1_{\{Q_j(t)>0\}}$, Lebesgue almost every $t \ge 0$. Let $t > 0$ be such that $Q_j(t) > 0$. Then, because of the continuity of Q_j, there exists a time interval (a, b) such that $t \in (a, b)$ and $Q_j(s) > 0$ for all $s \in (a, b)$; and (77) implies that $\kappa_j(s) = p_j\delta(s)$ for almost every $s \in (a, b)$. Next consider the set $Z = \{t \ge 0 : Q_j(t) = 0\}$, and let $t \in Z$ be a time at which derivatives of Q_j, E_j, R_j, and K_j exist and are equal to their densities (which holds for Lebesgue almost every $s \in Z$ because of the absolute continuity of Q_j, E_j, R_j, and K_j). From (45), (39), and (42), $Q_j(t) = Q_j(0) + E_j(t) - R_j(t) - K_j(t)$ and from (40) and $Q_j(t) = 0$, $\frac{d}{dt}R_j(t) = 0$. So

$$\frac{d}{dt}Q_j(t) = \lambda_j(t) - \kappa_j(t).$$

Because Q_j is nonnegative and $t \in Z$, $Q_j(t)$ must be a minimum, so $\frac{d}{dt}Q_j(t) = 0$, which implies $\kappa_j(t) = \lambda_j(t) \le p_j\delta(t)$. This, together with (77) implies that $\kappa_j(s) = \lambda_j(s) \le p_j\delta(s)$ for Lebesgue almost every s in Z. □

Remark 3. The proof of Lemma 11 suggests that $\lambda_j(\cdot) \wedge p_j\delta(\cdot)$ can be replaced with $\lambda_j(\cdot)$ in (79). However, we require $\lambda_j(\cdot) \wedge p_j\delta(\cdot)$ in (79) in order for the dynamics to be uniquely determined (in which case, either (78)–(79) or (77)–(78) can be used to specify the WRBS policy equations when E_j is absolutely continuous for each $j \in \mathbb{J}$).

5.2.2. Uniqueness of WRBS Fluid Model Solutions. In this section, we state and prove a uniqueness result concerning fluid model solutions. Specifically we prove the following.

Assumption 6. *For each, $j \in \mathbb{J}$, h_j^r is bounded.*

Theorem 3. *Suppose that E is an arrival function, $p \in \mathcal{P}$, and Assumption 6 holds. Let (X, ν, η) and $(\tilde{X}, \tilde{\nu}, \tilde{\eta})$ be fluid model solutions for E and the WRBS policy p such that $(X(0), \nu(0), \eta(0)) = (\tilde{X}(0), \tilde{\nu}(0), \tilde{\eta}(0))$. Then $(X, \nu, \eta) = (\tilde{X}, \tilde{\nu}, \tilde{\eta})$.*

The overall approach to proving Theorem 3 is similar to the approach used to prove uniqueness in Atar et al. [4] for static priority. In Atar et al. [4], they introduce a Skorokhod problem. Certain functionals of fluid model solutions operating under a fluid analog of static priority solve this Skorokhod problem. Uniqueness and Lipschitz continuity properties of solutions to this Skorokhod problem are leveraged to prove uniqueness of fluid model solutions operating under a fluid analog of static priority. It is not surprising that the same functionals of WRBS fluid model solutions do not necessarily solve the Skorokhod problem in Atar et al. [4] corresponding to static priority.

In order to prove the uniqueness theorem for WRBS, we specify a different Skorokhod problem (see Definition 7). Then we state and prove suitable uniqueness and Lipschitz continuity properties satisfied by solutions of this Skorokhod problem (see Lemma 12). Next, we observe that certain functionals of WRBS fluid model solutions solve this Skorokhod problem. Upon applying the Lipshitz continuity properties to WRBS fluid model solutions, we obtain Lipschitz continuity properties for WRBS fluid model solutions (see Lemma 13 and its proof). Once this is accomplished, the proof of Theorem 3 follows similarly to the proof of Atar et al. [4, theorem 3.1] upon using Lemma 13 in place of Atar et al. [4, proposition 3.3]. This part is straightforward to verify, and so the details of this are omitted.

That Assumption 6 holds is a hypothesis of both Atar et al. [4, theorem 3.1] and Theorem 3. The uniqueness proof in the multiclass case seems to require a fundamentally different approach than that given in Kaspi and Ramanan [12] for the single-class case, which does not require Assumption 6.

To begin, suppose that E is an arrival function and $p \in \mathcal{P}$. Let (X, ν, η) be a fluid model solution for E and the WRBS policy p. For $t \geq 0$, define

$$\check{K}_\Sigma(t) = \sum_{i=1}^{J-1} K_i(t) = K_\Sigma(t) - K_J(t).$$

Note that K_J is not part of the summation $\check{K}_\Sigma$. Then, by (42) and (43), for $t \geq 0$,

$$\check{K}_\Sigma(t) + K_J(t) = K_\Sigma(t) = B_\Sigma(t) + D_\Sigma(t) - B_\Sigma(0) = I(0) + D_\Sigma(t) - I(t).$$

Hence, for all $t \geq 0$,

$$K_J(t) + I(t) = I(0) + D_\Sigma(t) - \check{K}_\Sigma(t).$$

Further, by (45), (39), and (42), for $j \in \mathbb{J}$ and $t \geq 0$,

$$Q_j(t) = Q_j(0) + E_j(t) - R_j(t) - K_j(t).$$

Combining the two previous displayed equations yields that, for each $t \geq 0$,

$$\begin{aligned} Q_j(t) &= Q_j(0) + E_j(t) - R_j(t) - K_j(t), && \text{for } 1 \leq j < J, \\ Q_J(t) - I(t) &= Q_J(0) - I(0) + E_J(t) - R_J(t) - D_\Sigma(t) + \check{K}_\Sigma(t). \end{aligned}$$

For $t \geq 0$, set $\mathcal{Q}_j(t) = Q_j(t)$ for $1 \leq j < J$ and $\mathcal{Q}_J(t) = Q_J(t) - I(t)$. Then, for each $t \geq 0$,

$$\mathcal{Q}_j(t) = \mathcal{Q}_j(0) + E_j(t) - R_j(t) - p_j D_\Sigma(t) + \left(p_j D_\Sigma(t) - K_j(t)\right), \qquad \text{for } 1 \leq j < J, \tag{80}$$

$$\mathcal{Q}_J(t) = \mathcal{Q}_J(0) + E_J(t) - R_J(t) - p_J D_\Sigma(t) - \sum_{i=1}^{J-1} \left(p_i D_\Sigma(t) - K_i(t)\right). \tag{81}$$

By (78), $\mathcal{Q}^+ = Q$ and $\mathcal{Q}_J^- = I$. Therefore, uniquely solving these equations for $\mathcal{Q}$, uniquely specifies Q and I. Secondly, for each $1 \leq j < J$, it follows from (77) that $p_j D_\Sigma - K_j$ is nondecreasing and increases only when Q_j is zero.

This motivates the consideration of a Skorokhod problem as follows. The domain of this Skorokhod problem is $\mathbb{H}_+ = \mathbb{R}_+^{J-1} \times \mathbb{R}$. The reflection matrix $M \in \mathbb{R}^{J \times J-1}$ is given by

$$M = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -1 & -1 & -1 & \cdots & -1 \end{bmatrix}.$$

In particular, for $1 \le j < J$, the j^{th} column of M is denoted by $M_j \in \mathbb{R}^J$ and is given by $M_{j,j} = 1$, $M_{j,J} = -1$, and $M_{j,i} = 0$ for $1 \le i < J$ and $i \ne j$.

Definition 7 (SP ($\mathbb{H}_+$)). Let $\zeta \in \mathbf{D}\left(\mathbb{R}^J\right)$. Then (γ, υ), where $\gamma \in \mathbf{D}\left(\mathbb{R}_+^{J-1} \times \mathbb{R}\right)$ and $\upsilon \in \mathbf{D}\left(\mathbb{R}_+^{J-1}\right)$, are said to solve $\mathbf{SP}(\mathbb{H}_+)$ for ζ if

1. $\gamma = \zeta + M\upsilon$,
2. $\gamma(t) \in \mathbb{H}_+$ for all $t \ge 0$,
3. υ_j is nondecreasing and $\int_0^\infty 1_{\{\gamma_j(s)>0\}} d\upsilon_j(s) = 0$ for all $1 \le j < J$.

Lemma 12. *Given $\zeta \in \mathbf{D}\left(\mathbb{R}^J\right)$, $\boldsymbol{SP}(\mathbb{H}_+)$ has a unique solution (γ, υ). Furthermore, for any $\zeta, \tilde{\zeta} \in \mathbf{D}\left(\mathbb{R}^J\right)$ with respective unique solutions (γ, υ) and $(\tilde{\gamma}, \tilde{\upsilon})$ to $\boldsymbol{SP}(\mathbb{H}_+)$ and for any $t \ge 0$,*

$$\sum_{j=1}^{J-1} \|\upsilon_j - \tilde{\upsilon}_j\|_t \le \sum_{j=1}^{J-1} \|\zeta_j - \tilde{\zeta}_j\|_t, \tag{82}$$

$$\sum_{j=1}^{J} \|\gamma_j - \tilde{\gamma}_j\|_t \le 3 \sum_{j=1}^{J} \|\zeta_j - \tilde{\zeta}_j\|_t. \tag{83}$$

Proof. For each $1 \le j < J$, observe that $\gamma_j = \zeta_j + \upsilon_j$, where $\gamma_j \ge 0$ and υ_j is nondecreasing and increases only when γ_j is zero. For each $1 \le j < J$, this is a one-dimensional Skorokhod problem with unique solution

$$\upsilon_j(t) = \left(\sup_{0 \le s \le t} -\zeta_j(s) \right)^+.$$

Therefore, for $1 \le j < J$, υ_j is uniquely determined by ζ_j and consequently so is γ_j. Because $\gamma_J = \zeta_J - \sum_{j=1}^{J-1} \upsilon_j$ is uniquely determined by ζ_J and υ, uniqueness follows.

Given $\zeta, \tilde{\zeta} \in \mathbf{D}\left(\mathbb{R}^J\right)$ with respective unique solutions (γ, υ) and $(\tilde{\gamma}, \tilde{\upsilon})$, by the properties of the supremum function (e.g., see Whitt [21, lemma 13.4.1]), for each $1 \le j < J$ and $t \ge 0$, $\|\upsilon_j - \tilde{\upsilon}_j\|_t \le \|\zeta_j - \tilde{\zeta}_j\|_t$. Then, for each $1 \le j < J$ and $t \ge 0$, $\|\gamma_j - \tilde{\gamma}_j\|_t \le 2\|\zeta_j - \tilde{\zeta}_j\|_t$. Hence, for each $t \ge 0$,

$$\|\gamma_J - \tilde{\gamma}_J\|_t \le \|\zeta_J - \tilde{\zeta}_J\|_t + \sum_{j=1}^{J-1} \|\upsilon_j - \tilde{\upsilon}_j\|_t \le \sum_{j=1}^{J} \|\zeta_j - \tilde{\zeta}_j\|_t.$$

Therefore, (82) and (83) hold. □

In light of the existence and uniqueness of solutions to $\mathbf{SP}(\mathbb{H}_+)$, we make the following definition.

Definition 8. Let $\Gamma : \mathbf{D}\left(\mathbb{R}^J\right) \to \mathbf{D}\left(\mathbb{R}_+^{J-1} \times \mathbb{R}\right)$ and $\Upsilon : \mathbf{D}\left(\mathbb{R}^J\right) \to \mathbf{D}\left(\mathbb{R}_+^{J-1}\right)$ be given by $(\Gamma(\zeta), \Upsilon(\zeta)) = (\gamma, \upsilon)$, where (γ, υ) uniquely solves $\mathbf{SP}(\mathbb{H}_+)$ for $\zeta \in \mathbf{D}\left(\mathbb{R}^J\right)$.

Suppose that E is an arrival function and $p \in \mathcal{P}$. Let (X, ν, η) be a fluid model solution for E and the WRBS policy p. For $t \ge 0$, let

$$Z(t) = \mathcal{Q}(0) + E(t) - R(t) - pD_\Sigma(t). \tag{84}$$

Also, for $1 \le j < J$ and $t \ge 0$, let

$$U_j(t) = p_j D_\Sigma(t) - K_j(t). \tag{85}$$

Then by (80), (81), and the remarks immediately below those equations $\Gamma(Z) = \mathcal{Q}$ and $\Upsilon(Z) = U$, that is, $(\mathcal{Q}, U)$ uniquely solves $\mathbf{SP}(\mathbb{H}_+)$ for Z.

Given an arrival function E, $p \in \mathcal{P}$, and two fluid model solutions (X,ν,η) and $(\tilde{X},\tilde{\nu},\tilde{\eta})$ for E and the WRBS policy p, for $H = X, B, Q, D, K, R, I, \mathcal{Q}$, or Z and $\tilde{H} = \tilde{X}, \tilde{B}, \tilde{Q}, \tilde{D}, \tilde{K}, \tilde{R}, \tilde{I}, \tilde{\mathcal{Q}}$, or $\tilde{Z}$, let $\Delta H = H - \tilde{H}$. Observe that for all $t \geq 0$,

$$\|\Delta Q\|_t = \|\Delta \mathcal{Q}^+\|_t \leq \|\Delta \mathcal{Q}\|_t \tag{86}$$

Thus, similarly to Atar et al. [4, proposition 3.3], we have the following lemma:

Lemma 13. *Let E be an arrival function and $p \in \mathcal{P}$. Suppose that (X,ν,η) and $(\tilde{X},\tilde{\nu},\tilde{\eta})$ are fluid model solutions for E and the WRBS policy p such that $(X(0),\nu(0),\eta(0)) = (\tilde{X}(0),\tilde{\nu}(0),\tilde{\eta}(0))$. Then, there exists $c > 0$ such that for each $t \geq 0$, $\|\Delta Q\|_t \leq c(\|\Delta R\|_t + \|\Delta D\|_t)$.*

Proof. We have $(\Gamma(Z),\Upsilon(Z)) = (\mathcal{Q},U)$ and $(\Gamma(\tilde{Z}),\Upsilon(\tilde{Z})) = (\tilde{\mathcal{Q}},\tilde{U})$. By (86) and Lemma 12, for each $t \geq 0$, $\|\Delta Q\|_t \leq \|\Delta \mathcal{Q}\|_t \leq 3\|\Delta Z\|_t = 3(\|\Delta R\|_t + \|\Delta D\|_t)$. Thus, the result holds. □

Proof of Theorem 3. The result follows similarly to the proof of Atar et al. [4, theorem 3.1] by using Lemma 13 in place of Atar et al. [4, proposition 3.3]. The details of this are omitted. □

5.2.3. Invariant States for Weighted Random Buffer Selection. For time homogeneous systems where $E = \Lambda$ for some $\lambda \in (0,\infty)^J$, the set of invariant states for the (nonpolicy specific) fluid model was discussed in Section 3.3. By Proposition 1, $\mathbb{B}(\lambda)$ given by (53) is in one-to-one correspondence with the set of (nonpolicy-specific) invariant states for the arrival function Λ; for $b \in \mathbb{B}(\lambda)$, b_j corresponds to the proportion of server effort to be devoted to class j for the fluid model solution that is identically equal to the corresponding invariant state. The next result specifies the subset of $\mathbb{B}(\lambda)$ that is in one-to-one correspondence with the set of invariant states for WRBS policies.

Theorem 4. *Suppose that $\lambda \in (0,\infty)^J$ and that for each $j \in \mathbb{J}$, $m_j^s = \int_0^{H_j^s} 1 - G_j^s(x)dx < \infty$, $m_j^r = \int_0^{H_j^r} 1 - G_j^r(x)dx < \infty$ and G_j^r is strictly increasing with inverse $(G_j^r)^{-1}$, where by convention $(G_j^r)^{-1}(1) = H_j^r$. Let $\mathbb{B}(\lambda)$ be as in Proposition 1, recall that $\rho_J = \lambda_J m_J^s$ and define*

$$\mathbb{B}_J(\lambda) = \left\{ b \in \mathbb{B}(\lambda) : b_J = \min\left(\rho_J, 1 - \sum_{k=1}^{J-1} b_k \right) \right\}.$$

Given $b \in \mathbb{B}_J(\lambda)$, let $p(b) \in \mathcal{P}$ be given by

$$p_j(b) = \frac{b_j/m_j^s}{\sum_{k=1}^J b_k/m_k^s}, \qquad \text{for each } j \in \mathbb{J}. \tag{87}$$

For $b \in \mathbb{B}_J(\lambda)$, (X^*,ν^*,η^*) defined as in Proposition 1 is an invariant state for the arrival function Λ and WRBS policy $p(b)$. Conversely, given $p \in \mathcal{P}$ and an invariant state (X^*,ν^*,η^*) for the arrival function Λ and WRBS policy p, then $B^* \in \mathbb{B}_J(\lambda)$.

Before proceeding with the proof, we make an observation. For $\lambda \in (0,\infty)^J$, (87) in Theorem 4 specifies how to select the weights p_j for $j \in \mathbb{J}$ to achieve any allocation of server effort $b \in \mathbb{B}_J(\lambda)$ feasible by a WRBS policy. For the reader interested in nonidling policies, the corresponding allocations of server effort achievable by some WRBS invariant state for the arrival function Λ are the same as those for the (nonpolicy-specific) invariant states, for the arrival function Λ and given by

$$\left\{ b \in \mathbb{B}_J(\lambda) : \sum_{j=1}^J b_j = \min\left(1, \sum_{j=1}^J \lambda_j m_j^s\right) \right\} = \left\{ b \in \mathbb{B}(\lambda) : \sum_{j=1}^J b_j = \min\left(1, \sum_{j=1}^J \lambda_j m_j^s\right) \right\}.$$

Proof of Theorem 4. Suppose that $b \in \mathbb{B}_J(\lambda)$. Then $b \in \mathbb{B}(\lambda)$ and (X^*,ν^*,η^*) as defined in Proposition 1 is an invariant state for the arrival function Λ. Let $p(b)$ be given by (87) and let (X,ν,η) denote the function that is identically equal to (X^*,ν^*,η^*). Then (X,ν,η) is a fluid model solution for the arrival function Λ and it suffices to show that (X,ν,η) satisfies (77) and (78) in order to conclude that (X,ν,η) is a fluid model solution for the arrival function Λ and WRBS policy $p(b)$. By (41) and (42), $K_j(t) = D_j(t) = \frac{b_j}{m_j^s} t$ for all $t \geq 0$ and $j \in \mathbb{J}$. Thus, $D_\Sigma(t) = \left(\sum_{k=1}^J \frac{b_k}{m_k^s} \right) t$ for all $t \geq 0$. Hence, $K_j(t) = p_j(b) D_\Sigma(t)$ for all $t \geq 0$ and $j \in \mathbb{J}$ and (77) holds for $p = p(b)$. Also, because $b \in \mathbb{B}_J(\lambda)$ and $B_j(t) =$

b_j for all $t \geq 0$ and $j \in \mathbb{J}$, we have

$$I(t) = 1 - \sum_{k=1}^{J-1} b_k - b_J = \begin{cases} 1 - \sum_{k=1}^{J-1} b_k - \rho_J, & \text{if } b_J = \rho_J, \\ 0, & \text{if } b_J = 1 - \sum_{k=1}^{J-1} b_k. \end{cases}$$

Because $Q_J(t) = q_J(b_J)$ for all $t \geq 0$, where $q_J(b_J)$ is given by (54) and $q_J(\rho_J) = 0$, (78) holds.

Conversely, suppose that (X^*, ν^*, η^*) is an invariant state for the arrival function Λ and some WRBS policy $p \in \mathcal{P}$. Then (X^*, ν^*, η^*) is an invariant state for the arrival function Λ and (X^*, ν^*, η^*) are as given in Proposition 1 with $b = B^*$, where $B_j^* = \langle 1, \nu_j^* \rangle$ for all $j \in \mathbb{J}$. Let (X, ν, η) denote the fluid model solution for the arrival function Λ and WRBS policy p that is identically equal to (X^*, ν^*, η^*). Because of Proposition 1, $b \in \mathbb{B}(\lambda)$ and we must show that $b \in \mathbb{B}_J(\lambda)$. This follows from (78) because $B_j(t) = b_j$ for all $t \geq 0$ and $j \in \mathbb{J}$, $Q_J(t) = q_J(b_J)$ for all $t \geq 0$ and $q_J(\rho_J) = 0$. □

5.3. A Fluid Limit Theorem for WRBS Policies

Theorem 5. *Suppose that $p \in \mathcal{P}$ and $\{\bar{Y}^N\}_{N \in \mathbb{N}}$ is a sequence of fluid scaled state processes for WRBS policy p. Further suppose that Assumptions 4 and 5 hold and that (X, ν, η) is a distributional limit point of $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}_{N \in \mathbb{N}}$. Then (X, ν, η) is almost surely a fluid model solution for E and the WRBS policy p such that $(X(0), \nu(0), \eta(0)) = (X^0, \nu^0, \eta^0)$. In addition, (X, ν, η) and the auxiliary functions are continuous and $\eta_j(t)$ does not charge points for all $t \geq 0$ and $j \in \mathbb{J}$, almost surely. Moreover, if Assumption 6 also holds, then the law of (X, ν, η) is unique and $(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N) \Rightarrow (X, \nu, \eta)$ as $N \to \infty$.*

Proof. Because $\{(X^N, \nu^N, \eta^N)\}$ satisfies (55), Theorem 1 implies that $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}$ is tight. Suppose that $\{N'\}$ is a subsequence of $\{N\}$ such that $(\bar{X}^{N'}, \bar{\nu}^{N'}, \bar{\eta}^{N'}) \Rightarrow (X, \nu, \eta)$ as $N' \to \infty$. By Theorem 1, (X, ν, η) is, almost surely, a fluid model solution with initial value $(X(0), \nu(0), \eta(0)) = (X^0, \nu^0, \eta^0)$ that is continuous and has continuous auxiliary functions and such that $\eta_j(t)$ does not charge points for any $t \geq 0$ and $j \in \mathbb{J}$. It remains to show that the WRBS policy-specific Equations (77) and (78) hold.

Let $\{\bar{V}^N\}_{N \in \mathbb{N}}$ be as defined in Theorem 2 and suppose that $\tilde{V}$ is a limit point of $\{\bar{V}^{N'}\}$. Then, because Theorem 2 implies that $\{\bar{V}^N\}_{N \in \mathbb{N}}$ is tight, there exists a further subsequence $\{\bar{V}^{N''}\}$ such that $\{\bar{V}^{N''}\}$ converges to $\tilde{V}$. Because $\{(X^{N''}, \nu^{N''}, \eta^{N''})\}$ is a subsequence of $\{(X^{N'}, \nu^{N'}, \eta^{N'})\}$, it follows that $(\tilde{X}, \tilde{\nu}, \tilde{\eta})$ is equal in distribution to (X, ν, η). Let $V = (E, X, \nu, \eta, \langle 1, \eta \rangle, B, Q, R, D, K, I, D_{\sum}, \mathcal{D}, \mathcal{A}_s, \mathcal{S}, \mathcal{A}_r)$, where E is the limiting arrival process, (B, Q, R, D, K, I) are the auxiliary functions for (X, ν, η), $D_{\sum} = \sum_{j=1}^{J} D_j$ and $(\mathcal{D}, \mathcal{A}_s, \mathcal{S}, \mathcal{A}_r)$ are defined as in Lemma 10(2) using ν and η. Then V is equal in distribution to $\tilde{V}$; by invoking the Skorokhod representation theorem, without loss of generality, we may assume that $\{\bar{V}^{N''}\}$ converges to V almost surely. That Q_J and I satisfy (78) follows from this and (75).

It remains to verify that (77) holds. To ease the notation in what follows, we henceforth assume that $\{\bar{V}^N\}$ converges to V almost surely by relabeling the subsequence if necessary. We begin by showing that certain functions of the prelimit processes solve $\mathbf{SP}(\mathbb{H}_+)$. For each $N \in \mathbb{N}$ and $t \geq 0$, let $\mathcal{Q}_j^N(t) = Q_j^N(t)$ if $1 \leq j < J$ and $\mathcal{Q}_J^N(t) = Q_J^N(t) - I^N(t)$. Then, for each $N \in \mathbb{N}$ and $t \geq 0$,

$$\bar{\mathcal{Q}}_j^N(t) = \bar{\mathcal{Q}}_j^N(0) + \bar{E}_j^N(t) - \bar{R}_j^N(t) - \bar{K}_j^N(t), \qquad \text{for } 1 \leq j < J,$$

$$\bar{\mathcal{Q}}_J^N(t) = \bar{\mathcal{Q}}_J^N(0) + \bar{E}_J^N(t) - \bar{R}_J^N(t) + \sum_{j=1}^{J-1} \bar{K}_j^N(t) - \bar{D}_{\Sigma}^N(t).$$

The first of the two equations above is simply (20) with fluid scaling. To obtain the second, sum (19) over $j \in \mathbb{J}$; solve for K_J^N; use (16) to replace B_{Σ}^N with $N - I^N$ and $B_{\Sigma}^N(0)$ with $N - I^N(0)$; substitute the resulting expression for

K_J^N into (20) with $j = J$; simplify and apply fluid scaling. For each $N \in \mathbb{N}$, $j \in \mathbb{J}$, and $t \geq 0$, let

$$\bar{U}_j^N(t) = \int_0^t 1_{\left\{j^*_{D_\Sigma^N(u)} = j, Q_j^N(u-) = 0\right\}} d\bar{D}_\Sigma^N(u),$$

$$\bar{Z}_j^N(t) = \bar{\mathcal{Q}}_j^N(0) + \bar{E}_j^N(t) - \bar{R}_j^N(t) - \int_0^t 1_{\left\{j^*_{D_\Sigma^N(u)} = j\right\}} d\bar{D}_\Sigma^N(u).$$

Then, by (74), for $N \in \mathbb{N}$, $1 \leq j < J$, and $t \geq 0$,

$$\bar{K}_j^N(t) = \int_0^t 1_{\left\{j^*_{D_\Sigma^N(u)} = j\right\}} d\bar{D}_\Sigma^N(u) - \bar{U}_j^N(t).$$

It follows that, for $N \in \mathbb{N}$ and $t \geq 0$,

$$\bar{\mathcal{Q}}_j^N(t) = \bar{Z}_j^N(t) + \bar{U}_j^N(t), \quad \text{for } 1 \leq j < J, \text{ and} \quad \bar{\mathcal{Q}}_J^N(t) = \bar{Z}_J^N(t) - \sum_{j=1}^{J-1} \bar{U}_j^N(t).$$

Note that for each $N \in \mathbb{N}$ and $1 \leq j < J$, $\bar{U}_j^N$ is nondecreasing. In addition, for each $N \in \mathbb{N}$ and $1 \leq j < J$,

$$\int_0^\infty 1_{\left\{\bar{Q}_j^N(u) > 0\right\}} d\bar{U}_j^N(u) = \int_0^\infty 1_{\left\{\bar{Q}_j^N(u) > 0\right\}} 1_{\left\{j^*_{D_\Sigma^N(u)} = j, Q_j^N(u-) = 0\right\}} d\bar{D}_\Sigma^N(u) = 0,$$

almost surely, because if u is a departure time, then $Q_j^N(u-) = 0$ implies $Q_j^N(u) = 0$. Hence, for each $N \in \mathbb{N}$, $(\bar{\mathcal{Q}}^N, \bar{U}^N)$ uniquely solves $\mathbf{SP}(\mathbb{H}_+)$ for $\bar{Z}^N$, that is, $(\bar{\mathcal{Q}}^N, \bar{U}^N) = (\Gamma(\bar{Z}^N), \Upsilon(\bar{Z}^N))$.

For $j \in \mathbb{J}$ and $t \geq 0$, let

$$\varepsilon_j(t) = \sum_{i=1}^{\lfloor t \rfloor} \left(1_{\{j_i^* = j\}} - p_j\right) \text{ and } \bar{\varepsilon}_j^N(t) = \frac{\varepsilon_j(Nt)}{N}.$$

Then, for all $j \in \mathbb{J}$ and $t \geq 0$,

$$\bar{Z}_j^N(t) = \bar{\mathcal{Q}}_j^N(0) + \bar{E}_j^N(t) - \bar{R}_j^N(t) - \left(\bar{\varepsilon}_j N^\circ \bar{D}_\Sigma^N\right)(t) - p_j \bar{D}_\Sigma^N(t), \tag{88}$$

$$\bar{K}_j^N(t) = \left(\bar{\varepsilon}_j N^\circ \bar{D}_\Sigma^N\right)(t) + p_j \bar{D}_\Sigma^N(t) - \bar{U}_j^N(t). \tag{89}$$

Let $\mathcal{Q}_j = Q_j$ for $1 \leq j < J$ and $\mathcal{Q}_J = Q_J - I_J$ and define Z as in (84). Then, because $\lim_{N\to\infty} \bar{V}^N = V$ almost surely, $\lim_{N\to\infty} \bar{\mathcal{Q}}^N = \mathcal{Q}$ almost surely. In addition, because of the functional strong law of large numbers, $\lim_{N\to\infty} \bar{\varepsilon}_j^N = 0$ almost surely for each $j \in \mathbb{J}$. This together with $\lim_{N\to\infty} \bar{V}^N = V$ almost surely implies that $\lim_{N\to\infty} \bar{\varepsilon}_j^N \circ \bar{D}_\Sigma^N = 0$ for each $j \in \mathbb{J}$. The above together with (88) implies that $\lim_{N\to\infty} \bar{Z}^N = Z$ almost surely. Then, by the continuous mapping theorem, almost surely,

$$\lim_{N\to\infty} (\bar{\mathcal{Q}}^N, \bar{U}^N) = \lim_{N\to\infty} (\Gamma(\bar{Z}^N), \Upsilon(\bar{Z}^N)) = (\Gamma(Z), \Upsilon(Z)).$$

Hence, $\mathcal{Q} = \Gamma(Z)$. Set $U = \Upsilon(Z)$. Then, for each $1 \leq j < J$, U_j is a nonnegative, nondecreasing function that can only increase at times t when $\mathcal{Q}_j(t) = Q_j(t) = 0$. Also, because of (89), for each $1 \leq j < J$, $\lim_{N\to\infty} \bar{K}_j^N = p_j D_\Sigma - U_j$ almost surely. Therefore, for each $1 \leq j < J$,

$$K_j = p_j D_\Sigma - U_j, \tag{90}$$

and so for each $1 \leq j < J$ and $0 \leq s < t < \infty$,

$$K_j(t) - K_j(s) \leq \int_s^t p_j dD_\Sigma(u). \tag{91}$$

The potential times of increase of U_j and (90) imply that for all $1 \leq j < J$ and $0 \leq s < t < \infty$,

$$U_j(t) - U_j(s) = \int_s^t 1_{\{Q_j(u)=0\}} dU_j(u) = \int_s^t 1_{\{Q_j(u)=0\}} \left(p_j dD_\Sigma(u) - dK_j(u)\right).$$

For each $1 \le j < J$, the function K_j is nonnegative and nondecreasing because $\bar{K}_j^N$ is for each $N \in \mathbb{N}$. Hence, for each $1 \le j < J$ and $0 \le s < t < \infty$,

$$U_j(t) - U_j(s) \le p_j \int_s^t 1_{\{Q_j(u)=0\}} dD_\Sigma(u). \tag{92}$$

Then, for all $1 \le j < J$ and $0 \le s < t < \infty$, (90) and (92),

$$K_j(t) - K_j(s) = \int_s^t p_j dD_\Sigma(t) - (U_j(t) - U_j(s)) \ge p_j \int_s^t 1_{\{Q_j(u)>0\}} dD_\Sigma(u). \tag{93}$$

The inequalities (91) and (93) show that (77) holds. Thus, the limit point (X, ν, η) is a fluid model solution for E and WRBS policy p.

Finally, if Assumption 6 holds, then Theorem 3 implies that the fluid model solutions of every convergent subsequence have the same law, which, together with the tightness of $\{(\bar{X}^N, \bar{\nu}^N, \bar{\eta}^N)\}$, completes the proof. □

Acknowledgments

The authors thank Kavita Ramanan for many helpful discussions and Yunan Liu for his very careful read and helpful comments.

Appendix. Verification of Tightness

Here we summarize how to extend various components of the tightness argument given in Kang and Ramanan [11] for the single-class, nonidling m$\mathbb{S}$ any server queue to the multiclass many-server queue with control setting.

A.1. Criteria for Tightness

In the real valued setting, we have Kurtz' criteria (see, e.g., Ethier and Kurtz [8, theorem 3.8.6 and remark 3.8.7]).

A.1.1. Kurtz' Criteria. A sequence of processes $\{H^N\}_{N\in\mathbb{N}}$ with sample paths in $\mathbf{D}(\mathbb{R})$ is relatively compact if the following two properties hold:

(K.1) For all rational $t \ge 0$, $\lim_{M\to\infty} \sup_N \mathbb{P}(|H^N(t)| > M) = 0$.

(K.2) For all rational $t > 0$, there exists $q > 0$ such that $\lim_{\varepsilon\to 0} \sup_N \mathbb{E}\big[|H^N(t+\varepsilon) - H^N(t)|^q\big] = 0$.

In the measure valued setting, we have Jakubowski's criteria (see, e.g., Jakubowski [10, theorem 4.6]).

A.1.2. Jakubowski's Criteria. A sequence $\{\pi^N\} \subset \mathbf{D}(\mathbf{M}(\mathbb{S}))$ of random elements defined on $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathbb{S}$ is a Polish space, is tight if and only if the following two conditions hold:

(J.1) For each $T > 0$ and $\varepsilon > 0$, there exists a compact set $\mathcal{K}_{T,\varepsilon} \subset \mathbf{M}(\mathbb{S})$ such that

$$\liminf_{N\to\infty} \mathbb{P}\Big(\pi^N(t) \in \mathcal{K}_{T,\varepsilon} \text{for all } t \in [0,T]\Big) \ge 1 - \varepsilon.$$

(J.2) There exists a family $\mathbb{F}$ of real-valued continuous functions on $\mathbf{M}(\mathbb{S})$ that separates points in $\mathbf{M}(\mathbb{S})$ and is closed under addition such that for every $F \in \mathbb{F}$, $\{F(\pi^N(u)) : u \in [0,\infty)\}_{N\in\mathbb{N}}$ is tight in $\mathbf{D}(\mathbb{R})$.

Remark A.1. Given $H \in (0,\infty]$, let

$$\mathbb{F}_1(H) = \{F : \exists\, f \in \mathbf{C}_c^1([0,H)) \text{ such that } \mathrm{F}(\mu) = \langle f, \mu\rangle\ \forall \mu \in \mathbf{M}[0,H)\},$$
$$\mathbb{F}_2(H) = \{F : \exists\, f \in \mathbf{C}_c^{1,1}([0,H)\times\mathbb{R}_+) \text{ such that } \mathrm{F}(\mu) = \langle f, \mu\rangle\ \forall \mu \in \mathbf{M}([0,H)\times\mathbb{R}_+)\}.$$

Then, $\mathbb{F}_1(H)$ and $\mathbb{F}_2(H)$ are families of real-valued continuous functions on $\mathbf{M}[0,H)$ and $\mathbf{M}([0,H)\times\mathbb{R}_+)$, respectively, that separate points and are closed under addition.

A.2. Preliminary Estimates for Verifying Tightness

The main goal of this section is to establish the next lemma, which is analogous to Kaspi and Ramanan [12, lemma 5.8] and used to verify tightness in the next subsection.

Lemma A.1. *Suppose that Assumptions 1 and 3 hold and for each $j \in \mathbb{J}$,*

$$\lim_{m\uparrow H_j^s} \sup_N \mathbb{E}\Big[\Big\langle 1_{(m,H_j^s)}, \bar{\nu}_j^N(0)\Big\rangle\Big] = 0 \quad \text{and} \quad \lim_{m\uparrow H_j^r} \sup_N \mathbb{E}\Big[\Big\langle 1_{(m,H_j^r)}, \bar{\eta}_j^N(0)\Big\rangle\Big] = 0. \tag{A.1}$$

In addition, suppose that for $j \in \mathbb{J}$ such that $H_j^s < \infty$ (resp. $H_j^r < \infty$),

$$\lim_{m\uparrow H_j^s} \sup_N \mathbb{E}\left[\left\langle 1_{[0,m)}(\cdot)\frac{1-G_j^s(m)}{1-G_j^s(\cdot)}, \bar{\nu}_j^N(0)\right\rangle\right] = 0 \tag{A.2}$$

$$\left(\text{resp. } \lim_{m\uparrow H_j^r} \sup_N \mathbb{E}\left[\left\langle 1_{[0,m)}(\cdot)\frac{1-G_j^r(m)}{1-G_j^r(\cdot)}, \bar{\eta}_j^N(0)\right\rangle\right] = 0\right). \tag{A.3}$$

Then, for each $j \in \mathbb{J}$ and $t \ge 0$, the following hold:

1. *We have*

$$\lim_{m\uparrow H_j^s}\sup_N \mathbb{E}\left[\int_0^t \left\langle 1_{(m,H_j^s)}h, \bar{\nu}_j^N(u)\right\rangle du\right] = 0 \qquad \text{and} \qquad \lim_{m\uparrow H_j^r}\sup_N \mathbb{E}\left[\int_0^t \left\langle 1_{(m,H_j^r)}h, \bar{\eta}_j^N(u)\right\rangle du\right] = 0.$$

2. *For* $\varphi \in \mathbf{C}_b([0,H_j^s)\times\mathbb{R}_+)$ *and* $\psi \in \mathbf{C}_b([0,H_j^r)\times\mathbb{R}_+)$,

$$\lim_{\delta\searrow 0}\limsup_{N\to\infty}\mathbb{E}\left[\bar{\mathcal{A}}_{s,j}^N(\varphi,t+\delta)-\bar{\mathcal{A}}_{s,j}^N(\varphi,t)\right]=0, \quad \lim_{\delta\searrow 0}\limsup_{N\to\infty}\mathbb{E}\left[\bar{\mathcal{D}}_j^N(\varphi,t+\delta)-\bar{\mathcal{D}}_j^N(\varphi,t)\right]=0,$$
$$\lim_{\delta\searrow 0}\limsup_{N\to\infty}\mathbb{E}\left[\bar{\mathcal{A}}_{r,j}^N(\psi,t+\delta)-\bar{\mathcal{A}}_{r,j}^N(\psi,t)\right]=0, \quad \lim_{\delta\searrow 0}\limsup_{N\to\infty}\mathbb{E}\left[\bar{\mathcal{S}}_j^N(\psi,t+\delta)-\bar{\mathcal{S}}_j^N(\psi,t)\right]=0,$$

3. *Given* $m < H_j^s$ *(resp.* $m < H_j^r$*) and a sequence* $\{\mathcal{H}_n\}_{n\in\mathbb{N}}$ *of Borel subsets of* $[0,m]$ *with Lebesgue measure converging to zero as* $n\to\infty$*, we have*

$$\lim_{n\to\infty}\limsup_{N\to\infty}\mathbb{E}\left[\sup_{0\le u\le t}\bar{\mathcal{A}}_{s,j}^N(1_{\mathcal{H}_n},u)\right]=0 \qquad \left(\text{resp.}\ \lim_{n\to\infty}\limsup_{N\to\infty}\mathbb{E}\left[\sup_{0\le u\le t}\bar{\mathcal{A}}_{r,j}^N(1_{\mathcal{H}_n},u)\right]=0\right).$$

The proof of Lemma A.1 is nearly identical to the proof of Kaspi and Ramanan [12, lemma 5.8] as given there, with the addition of class-level subscripts, where one follows the comments in Kang and Ramanan [11, remark 5.2] to address reneging. That proof uses Kang and Ramanan [12, lemma 5.6 and proposition 5.7]. To facilitate this, we state and verify analogs of those results applicable to the present setting.

First, we state and prove Lemma A.2, which is a version of Kaspi and Ramanan [12, lemma 5.6] adapted to the present multiclass setting with reneging and control.

Lemma A.2. *For* $N\in\mathbb{N}$, $t\ge 0$, *and* $j\in\mathbb{J}$,

$$\lim_{\delta\searrow 0}\mathbb{E}\left[\bar{D}_j^N(t+\delta)-\bar{D}_j^N(t)\right]=0, \tag{A.4}$$

$$\lim_{\delta\searrow 0}\mathbb{E}\left[\bar{S}_j^N(t+\delta)-\bar{S}_j^N(t)\right]=0. \tag{A.5}$$

In addition, for $N\in\mathbb{N}$, $t\ge 0$, $j\in\mathbb{J}$, $0<\delta<H_j^s$, *and* $m\in[0,H_j^s-\delta)$,

$$\bar{\mathcal{D}}_j^N(1_{[m+\delta,H_j^s)},t+\delta)-\bar{\mathcal{D}}_j^N(1_{[m+\delta,H_j^s)},t)\le\left\langle 1_{[m,H_j^s)},\bar{\nu}_j^N(t)\right\rangle. \tag{A.6}$$

Similarly, for $N\in\mathbb{N}$, $t\ge 0$, $j\in\mathbb{J}$, $0<\delta<H_j^r$, *and* $m\in[0,H_j^r-\delta)$,

$$\bar{\mathcal{S}}_j^N(1_{[m+\delta,H_j^r)},t+\delta)-\bar{\mathcal{S}}_j^N(1_{[m+\delta,H_j^r)},t)\le\left\langle 1_{[m,H_j^r)},\bar{\eta}_j^N(t)\right\rangle. \tag{A.7}$$

The proof of (A.4) is similar to the proof of the analogous statement in Kaspi and Ramanan [12, lemma 5.6] but requires an adaptation that we develop in the proof of Lemma A.2. For the single-class queue, the generalization to the case with reneging is discussed in Kang and Ramanan [11, remark 5.2]. Although the statements for the potential reneging process are exactly analogous to the statements for the departure process, the proofs of the two are different. Therefore, we have taken the opportunity to clarify that in the proof given here as well. Finally, we remark that in (A.6) and (A.7), we have omitted the conditional expected value and the renewal function term because they are not needed, which is also justified in the proof given below.

Proof of Lemma A.2. The proof of Kaspi and Ramanan [12, lemma 5.6] applies in the single-class case and requires one to consider a modified system that provides a bound on the total number of departures in a short time interval. A modification of this is needed to carry out the proof in the present setting. For this, fix $N\in\mathbb{N}$, $j\in\mathbb{J}$, and $t\in[0,T]$. In the multiclass setting, the original system and modified system have the same state at time t, except that in the modified system the class j queue length is taken to be infinity and any server that is idle at time t in the original system begins serving a class j customer at time t. The modified system evolves from time t in a nonidling fashion with priority to class j. For $i\in\mathbb{J}$, $x\in[0,H_i^s)$, and $\delta>0$, let $\tilde{D}_{j,i}^N(\delta\,|\,x)$ be the number of renewals in a delayed renewal process with interarrival distribution G_j^s and initial delay distribution $G_{i,x}^s$. Then, for $i\in\mathbb{J}$, $x\in[0,H_i^s)$, and $\delta>0$, $\mathbb{E}\left[\tilde{D}_{j,i}^N(\delta\,|\,x)\right]\le U_j^s(\delta)$, where U_j^s is the zero delay renewal function, which is given by $U_j^s(x)=\sum_{n=0}^\infty (G_j^s)^{(*n)}(x)$, for all $x\in\mathbb{R}_+$, where for $x\in\mathbb{R}_+$, $(G_j^s)^{(*0)}(x)=1$ and $(G_j^s)^{(*n)}(x)=\int_0^x G_j^s(x-y)g_j(y)dy$ for $n\in\mathbb{N}$. Hence, as in Kaspi and Ramanan [12, (5.30)], one obtains that for $\delta>0$,

$$\mathbb{E}\left[\bar{D}_j^N(t+\delta)-\bar{D}_j^N(t)\,|\,\mathcal{F}_t^N\right]\le\sum_{i=1}^J\left\langle\mathbb{E}[\tilde{D}_{j,i}^N(\delta\,|\,\cdot)],\bar{\nu}_i^N(t)\right\rangle+\bar{I}^N(t)\mathbb{E}[\tilde{D}_{j,j}^N(\delta\,|\,0)]\le U_j^s(\delta). \tag{A.8}$$

As in the proof of Kaspi and Ramanan [12, lemma 5.6], $\lim_{\delta\searrow 0}\mathbb{E}\left[\bar{D}_j^N(t+\delta)-\bar{D}_j^N(t)\,|\,\mathcal{F}_t^N\right]=0$ because for each $i\in\mathbb{J}$, $\lim_{\delta\searrow 0}\left\langle\mathbb{E}\left[\tilde{D}_{j,i}^N(\delta\,|\,\cdot)\right],\bar{\nu}_i^N(t)\right\rangle=0$ by bounded convergence because $\mathbb{E}\left[\tilde{D}_{j,i}(\delta|x)\right]\searrow 0$ as $\delta\searrow 0$ for all $x\in[0,H_i^s)$ and $U_j^s(\delta)$ is finite for

all $\delta > 0$ and nondecreasing. Then (A.4) follows by a second application of bounded convergence applied to $\mathbb{E}[\mathbb{E}[\bar{D}_j^N(t+\delta) - \bar{D}_j^N(t) \,|\, \mathcal{F}_t^N]]$ as $\delta \searrow 0$.

The proofs of (A.5), (A.6), and (A.7) do not require the use of a modified system. In particular, by considering separately customers that arrive to the original system during the time interval $(t, t+\delta]$ and customers in the original system at time t, we obtain that for $\delta > 0$,

$$\mathbb{E}\left[\bar{S}_j^N(t+\delta) - \bar{S}_j^N(t) | \mathcal{F}_t^N\right] \le \mathbb{E}\left[\bar{E}_j^N(t+\delta) - \bar{E}_j^N(t)\right] G_j^r(\delta) + \int_0^{H_j^r} \frac{G_j^r(x+\delta) - G_j^N(x)}{1 - G_j^N(x)} \bar{\eta}_j^N(t)(dx).$$

For each $x \in [0, H_j^r)$ and $\delta > 0$,

$$0 \le \frac{G_j^r(x+\delta) - G_j^N(x)}{1 - G_j^N(x)} \le 1 \quad \text{and} \quad \lim_{\delta \searrow 0} \frac{G_j^r(x+\delta) - G_j^N(x)}{1 - G_j^N(x)} = 0.$$

Then (A.5) also follows by bounded convergence.

To verify (A.6), for $0 < \delta < H_j^s$, and $m \in [0, H_j^s - \delta)$, note that any class j customer that departs in $(t, t+\delta]$ and is of age at least $m + \delta$ had to be in service and of age at least m at time t, which implies (A.6). Similarly, (A.7) holds because any customer with a waiting time at least $m + \delta$ that reneges during the time interval $(t, t+\delta]$ must have been waiting in system for at least m time units at time t. □

Next we state a version of Kaspi and Ramanan [12, proposition 5.7] applicable in the multiclass setting. This proposition is applicable because the hazard functions of the service time and reneging distributions are locally integrable. It is the other result used in the proof of Lemma A.1.

Proposition A.1. *Given $N \in \mathbb{N}$, $j \in \mathbb{J}$, $\ell \in \mathbf{L}^1_{\mathrm{loc}}[0, H_j^s)$, and $\varphi \in \mathbf{C}_b([0, H_j^s) \times \mathbb{R}_+)$, $\int_0^{\cdot} \left\langle \varphi(\cdot, u)\ell(\cdot), \bar{\nu}_j^N(u) \right\rangle du$ is well defined. Moreover, if $\ell \in \mathbf{L}^1[0, H_j^s)$, then for all $0 \le u \le t < \infty$,*

$$\left| \int_u^t \left\langle \varphi(\cdot, v)\ell(\cdot), \bar{\nu}_j^N(v) \right\rangle dv \right| \le \|\varphi\|_\infty \left(\bar{X}_j^N(0) + \bar{E}_j^N(t) \right) \sup_{0 \le v < H_j^s} \int_v^{(v+t-u) \wedge H_j^s} |\ell(x)| \, dx.$$

Similarly, given $N \in \mathbb{N}$, $j \in \mathbb{J}$, $\ell \in \mathbf{L}^1_{\mathrm{loc}}[0, H_j^r)$, and $\psi \in \mathbf{C}_b([0, H_j^r) \times \mathbb{R}_+)$, $\int_0^{\cdot} \langle \psi(\cdot, u)\ell(\cdot), \bar{\eta}_j^N(u) \rangle du$ is well defined. Moreover, if $\ell \in \mathbf{L}^1[0, H_j^r)$, then for all $0 \le u \le t < \infty$,

$$\left| \int_u^t \left\langle \psi(\cdot, v)\ell(\cdot), \bar{\eta}_j^N(v) \right\rangle dv \right| \le \|\psi\|_\infty \left(\left\langle 1, \bar{\eta}_j^N(0) \right\rangle + \bar{E}_j^N(t) \right) \sup_{0 \le v < H_j^r} \int_v^{(v+t-u) \wedge H_j^r} |\ell(x)| \, dx.$$

The proof of Proposition A.1 follows similarly to the proof of Kaspi and Ramanan [12, proposition 5.7] by adding class-level subscripts and/or applying analogous properties of the potential reneging measure. As such, we leave the verification as an exercise for the interested reader.

A.3. Compact Containment of the Measure Valued Processes

In order to demonstrate tightness of various measure valued processes, we verify that the conditions in Jakubowski's criteria hold under suitable assumptions (see Section A.1 for the statement of Jakubowski's criteria). The focus of this section is to verify the compact containment condition (J.1) of Jakubowski's criteria is satisfied for various measure value processes, which largely follows the arguments given in Kang and Ramanan [11] and Kaspi and Ramanan [12] for the single-class nonidling many-server queue.

Lemma A.3. *Suppose that Assumptions 1 and 3 hold. Then, for each $j \in \mathbb{J}$, (A.1), (A.2), and (A.3) hold. In addition, for each $j \in \mathbb{J}$, $\{\bar{\nu}_j^N\}_{N \in \mathbb{N}}$ and $\{\bar{\eta}_j^N\}_{N \in \mathbb{N}}$ satisfy (J.1) of Jakubowski's criteria.*

Lemma A.3 can be verified similarly to the arguments in the proofs of Kaspi and Ramanan [12, lemma 5.12] and Kang and Ramanan [11, lemma 6.6] by adding class-level subscripts, tracing the arguments given therein, and noting that none of those arguments rely on the specifics of the entry-into-service process. To see this, fix $j \in \mathbb{J}$. For $H_j^s = \infty$ (resp. $H_j^r = \infty$) that $\{\bar{\nu}_j^N\}_{N \in \mathbb{N}}$ (resp. $\{\bar{\eta}_j^N\}_{N \in \mathbb{N}}$) satisfies (J.1) of Jakubowski's criteria and that (A.1) holds can be established similarly to the proof of this given in the $M = \infty$ case of Kaspi and Ramanan [12, lemma 5.12] (resp. Kang and Ramanan [11, lemma 6.6]). Subsequently, in case $H_j^s < \infty$ (resp. $H_j^r < \infty$) that (A.1) holds can be established similarly to the proof given in the $M < \infty$ case of Kaspi and Ramanan [12, lemma 5.12] (resp. Kang and Ramanan [11, lemma 6.6]). Next, (A.2) (resp. (A.3)) can be established similarly to the proof given in the $M < \infty$ case of Kaspi and Ramanan [12, lemma 5.12] (resp. Kang and Ramanan [11, lemma 6.6]). Now that (A.1), (A.2), and (A.3) have been shown to hold, the fact that $\{\bar{\nu}_j^N\}_{N \in \mathbb{N}}$ and $\{\bar{\eta}_j^N\}_{N \in \mathbb{N}}$ satisfy (J.1) of Jakubowski's criteria in case $H_j^s < \infty$ (resp. $H_j^r < \infty$) can be established similarly to the conclusion of the $M < \infty$ cases of the proof of Kaspi and Ramanan [12, Lemma 5.12] (resp. Kaspi and Ramanan Kang and Ramanan [11, lemma 6.6]) by utilizing

Lemma 4 in place of Kaspi and Ramanan [12, corollary 5.5] and Kang and Ramanan [11, equation (5.3) of proposition 5.1] and Lemma A.1(1) in place of Kaspi and Ramanan [12, lemma 5.8(1)] and Kang and Ramanan [11, remark 5.2].

Lemma A.4. *Suppose that Assumptions 1 and 3 hold. Then, for each* $j \in \mathbb{J}$, $\{\bar{\mathcal{D}}_j^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{A}}_{s,j}^N\}_{N\in\mathbb{N}}$, $\{\bar{\mathcal{S}}_j^N\}_{N\in\mathbb{N}}$, *and* $\{\bar{\mathcal{A}}_{r,j}^N\}_{N\in\mathbb{N}}$ *satisfy* (J.1) *of Jakubowski's criteria.*

The proof of Lemma A.4 follows similarly to the arguments that (J.1) holds as given in the proof of Kaspi and Ramanan [12, lemma 5.13] by incorporating reneging as in Kang and Ramanan [11, lemma 6.7] and by adding class-level subscripts. In this regard, note that Lemma 4 and (57) of Lemma 5 can be used in place of the first assertion in Kaspi and Ramanan [12, lemma 5.6] and Kang and Ramanan [11, proposition 5.2] and Lemma A.1(1) in place of Kaspi and Ramanan [12, lemma 5.8(1)].

Endnotes

[1] Additionally, from Bassamboo and Randhawa [5], HL scheduling may not be optimal. Still, we focus on HL scheduling, as it is a common scheduling discipline.

[2] In Puha and Ward [19], continuity assumptions are made that seem to require adding the condition that K is continuous here. However, if (X, ν, η) is a constant fluid model solution for arrival function E, then because of (41) and (42), K_j is absolutely continuous for each $j \in \mathbb{J}$. So Proposition 1 holds without a priori requiring K to be continuous.

References

[1] Afeche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing Service Oper. Management* 15(3):423–443.
[2] Afeche P, Pavlin JM (2016) Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Sci.* 62(8):2412–2436.
[3] Atar R, Giat C, Shimkin N (2010) The $\frac{c\mu}{\theta}$ rule for many-server queues with abandonment. *Oper. Res.* 58(5):1427–1439.
[4] Atar R, Kaspi H, Shimkin N (2014) Fluid limits for many-server systems with reneging under a priority policy. *Math. Oper. Res.* 39(3): 672–696.
[5] Bassamboo A, Randhawa RS (2016) Scheduling homogeneous impatient customers. *Management Sci.* 62(7):2129–2147.
[6] Billingsley P (1999) *Convergence of Probability Measures*, 2nd ed. (John Wiley & Sons, Inc., New York).
[7] Dai JG, He S (2012) Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *J. Systems Sci. Systems Engrg.* 21:1–36.
[8] Ethier SN, Kurtz TG (1986) *Markov Processes: Characterization and Convergence* (Wiley, New York).
[9] Folland GB (1984) *Real Analysis: Modern Techniques and Their Applications* (John Wiley & Sons, New York).
[10] Jakubowski A (1986) On the Skorokhod topology. *Ann. Inst. Henri Poincare Probab. Statist.* 22(3):263–285.
[11] Kang W, Ramanan K (2010) Fluid limits of many-server queues with reneging. *Ann. Appl. Probab.* 20(6):2204–2260.
[12] Kaspi H, Ramanan K (2011) Law of large numbers limits for many-server queues. *Ann. Appl. Probab.* 21(1):33–114.
[13] Kim J, Ward AR (2013) Dynamic scheduling of a queue with multiple customer classes. *Queueing Systems* 75(2):339–384.
[14] Kim J, Randhawa RS, Ward AR (2018) Dynamic scheduling in a many-server multi-class system: The role of customer impatience in large systems. *Manufacturing Service Oper. Management* 20(2):285–301.
[15] Lee N (2008) A sufficient condition for stochastic stability of an Internet congestion control model in terms of fluid model stability. Unpublished PhD thesis, University of California, San Diego.
[16] Liu Y, Whitt W (2012) The $G_t/GI/s_t$ + GI many-server fluid queue. *Queueing Systems* 71:405–444.
[17] Long Z, Shimkin N, Zhang H, Zhang J (2019) Dynamic scheduling of multiclass many-server queues with abandonment: The generalized $c\mu/h$ rule. *Oper. Res.* 68(4):1218–1230.
[18] Pinedo ML (2012) *Scheduling: Theory, Algorithms, and Systems* (Springer Science & Business Media, Berlin).
[19] Puha AL, Ward AR (2019) Scheduling an overloaded multiclass many-server queue with impatient customers. Netessine S, ed. *Operations Research & Management Science in the Age of Analytics*, Tutorials in Operations Research (INFORMS, Catonsville, MD), 189–217.
[20] Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys Oper. Res. Management Sci.* 17(1):1–14.
[21] Whitt W (2002) *Stochastic Process Limits* (Springer, New York).
[22] Whitt W (2006) Fluid models for multiserver queues with abandonments. *Oper. Res.* 54(1):37–54.
[23] Wierman A (2007) Fairness and classifications. *ACM Performance Evaluation Rev.* 34(4):4–12.
[24] Zhan D, Ward AR (2019) Staffing, routing, and payment to trade off speed and quality in large service systems. *Oper. Res.* 67(6): 1738–1751.
[25] Zhang J (2013) Fluid models of many-server queues with abandonment. *Queueing Systems* 73(2):147–193.