Stochastic Systems
2013, Vol. 3, No. 1, 262–321
DOI: 10.1214/12-SSY085

FLUID LIMITS FOR OVERLOADED MULTICLASS FIFO SINGLE-SERVER QUEUES WITH GENERAL ABANDONMENT

By Otis B. Jennings and Amber L. Puha*

California State University San Marcos

We consider an overloaded multiclass nonidling first-in-first-out single-server queue with abandonment. The interarrival times, service times, and deadline times are sequences of independent and identically, but generally distributed random variables. In prior work, Jennings and Reed studied the workload process associated with this queue. Under mild conditions, they establish both a functional law of large numbers and a functional central limit theorem for this process. We build on that work here. For this, we consider a more detailed description of the system state given by K finite, nonnegative Borel measures on the nonnegative quadrant, one for each job class. For each time and job class, the associated measure has a unit atom associated with each job of that class in the system at the coordinates determined by what are referred to as the residual virtual sojourn time and residual patience time of that job. Under mild conditions, we prove a functional law of large numbers for this measure-valued state descriptor. This yields approximations for related processes such as the queue lengths and abandoning queue lengths. An interesting characteristic of these approximations is that they depend on the deadline distributions in their entirety.

1. Introduction. Here we consider a single-server queue fed by K arrival streams, each corresponding to a distinct job class. Upon arrival, each job declares its service time and deadline requirements. If a job doesn't enter service within the deadline time of its arrival, it abandons the queue before initiating service. Otherwise it remains in queue until it receives its full service time requirement. Nonabandoning jobs are served in a nonidling, first-in-first-out (FIFO) fashion. The queue is assumed to be overloaded, i.e., the offered load exceeds one. In addition, it is assumed that the interarrival times, service times, and deadline times are mutually independent sequences

Received October 2012.

*Research supported in part by UCLA-IPAM

AMS 2000 subject classifications: Primary 60B12, 60F17, 60K25; secondary 68M20, 90B22.

Keywords and phrases: Overloaded queue, abandonment, first-in-first-out, multiclass queue, measure-valued state descriptor, queue-length vector, fluid limits, fluid model, invariant states.

of independent and identically, but generally distributed random variables. One aim of this work is to provide approximations for functionals such as the queue-length process.

Queueing systems with abandonment are observed in many applications. Hence, it is a natural phenomenon to study. For FIFO single-server queues with abandonment, the earliest analysis focused on models with exponential abandonment [1], which is not an ideal modeling assumption. Thereafter, general abandonment distributions were considered in [3], which restricted to Markovian service and arrival processes. Shortly thereafter, stability conditions were determined in [2] without the Markovian restriction. In the last decade, there has been considerable progress with analyzing the critically loaded FIFO single-server queue with general abandonment [7, 19]. Here we focus on general abandonment for the multiclass overloaded FIFO single-server model described above.

In [11], Jennings and Reed study the workload process associated with this queue under the assumption that the abandonment distributions are continuous. Note that the workload process is well defined since each job declares its deadline upon arrival. In particular, the value of the workload process is increased by a job's service time at its arrival time if and only if the value of the workload process immediately before the job's arrival is strictly less than the job's deadline time. As usual, the value of the workload decreases at rate one while the workload is positive. Under mild conditions, they establish both a functional law of large numbers and a functional central limit theorem for this process.

In this paper, we build on that work. One outcome is to provide a fluid approximation for the vector-valued queue-length process. Because the abandonment distributions are not necessarily exponential, the queue-length vector together with the vector of residual interarrival times, class in service, and residual service time does not provide a Markovian description of the system state. Additional information about the residual deadline times is also needed. We track this information using a measure-valued state descriptor, which is defined precisely in Section 2.1. We give an informal description here. The state $\mathcal{Z}(t)$ of the system at time t consists of K finite, nonnegative Borel measures on $\mathbb{R}_+^2$, one for each class, where $\mathbb{R}_+$ denotes the nonnegative real numbers. Each measure consists entirely of unit atoms, one corresponding to each job of that class in the system. The coordinates of each atom are determine by two quantities associated with the job. The first coordinate is the residual virtual sojourn time of that job, which is the amount of work in the system (the cumulative residual service times of jobs that don't abandon) associated with this job and the jobs that arrived ahead of it. The second coordinate is the residual patience time. This is given by the job's deadline minus the time in system if it will abandon before entering service, and is given by the job's deadline plus its service time minus the time in system otherwise. Note that the atoms associated with jobs that will eventually be served are initially located in $\mathbb{R}_+^2$ above the line of slope one intersecting the origin, while the atoms associated with jobs that won't be served are initially placed on or below this line. With this description, new jobs arrive and the corresponding unit atoms are added at the appropriate coordinates of the system state. Further, each coordinate of each atom decreases at rate one until the unit atom reaches one of the coordinate axes and exits the system. Jobs associated with atoms that hit the vertical coordinate axis exit due to service completion. Jobs associated with atoms that ultimately hit the horizontal coordinate axis exit due to abandonment (see Figure 1).

Our choice of state descriptor is reminiscent of the one used by Gromoll, Robert, and Zwart [9] to analyze an overloaded processor sharing (PS) queue with abandonment. One distinction is that their work is for a single-class queue, and therefore has one coordinate rather than K. Another is that the first coordinate for the unit atom associated with a given job in the PS system is simply the job's residual service time. We use the residual virtual sojourn time to determine the first coordinate since it captures the order of arrivals, which is needed for FIFO. The evolution of the state descriptor in [9] is slightly more complicated than the one used here. In both cases, the residual patience times decrease at rate one, but in [9] the service times decrease at rate one over the number of jobs in the queue. Hence, the nice relationship with the line of slope one intersecting the origin present in the FIFO model is not present in the PS model. However, in both models it is true that hitting a coordinate axis corresponds to exiting the system with the vertical axis being associated with service completion and the horizontal axis being associated with abandonment. So this work provides an example of how the modeling framework developed for analyzing PS with abandonment can be adapted to yield an analysis in another abandonment setting.

Measure valued processes have been used rather extensively for modeling many server queues with and without abandonment (see [13, 14, 15, 16], and [17]). An important distinction between single-server and many server first-come-first-serve queues is that the later is not first-in-first-out. Therefore the measure valued descriptor and analysis given here is quite different from that found in the many server queues literature.

We develop a fluid approximation for this measure-valued state descriptor, which yields fluid approximations for the queue-length vector and other functionals of interest. Similarly to the approximations derived in [11] and [20], these approximations depend on the entire abandonment distribution of each job class. We begin by introducing an associated fluid model, which can be viewed as a formal law of large numbers limit of the stochastic system. Hence fluid model solutions are K-dimensional measure-valued functions with each coordinate taking values the space of finite, nonnegative Borel measures on $\mathbb{R}_+^2$ that satisfy an appropriate fluid model equation (see (26)). We analyze the behavior of fluid model solutions. Through this analysis we identify a nonlinear mapping from the fluid workload to the fluid queue-length vector (see (30)). Using the nature of fluid model solutions, this mapping is refined to yield the approximations for the number of jobs of each class in queue that will and will not abandon (see (31) and (32)). In a similar spirit, we obtain approximations for the number of jobs of each class in system of a certain age or older (see (33) and (34)). In addition, we characterize the invariant states for this fluid model (Theorem 3.3).

Next we justify interpreting fluid model solutions and functionals derived from them as first order approximations of prelimit functionals in the stochastic model by proving a fluid limit theorem (Theorem 3.2). It states that under mild assumptions the fluid scaled state descriptors for the stochastic system converge in distribution to measure-valued functions that are almost surely fluid model solutions. A basic element of our fluid limit result is the nature of the scaling employed. Consistent with the framework in [11], we accelerate both the arrival process and the service rates, while leaving the abandonment times unchanged. One should think of a concomitant speeding up of the server's processing rate to accommodate the increased customer demand; the content of the work for any particular customer is the same. We assume any given customer is unaffected by the increase in the number of fellow customers, but rather it is the time in queue that triggers abandonment. Hence, abandonment propensity does not require adjusting as the other system parameters increase.

The paper is organized as follows. We conclude this section with a listing of our notation. The formal stochastic and fluid models are given Section 2. Then, in Section 3, we present the main results (Theorems 3.1, 3.2 and 3.3) and several approximations derived from our fluid model. Theorems 3.1 and 3.3 are proved in Section 4. In the final two sections, we provide the proof of Theorem 3.2. The proof of tightness is presented in Section 5, while the characterization of fluid limit points as fluid model solutions is presented in Section 6. For this, we prove a functional law of large numbers for a sequence of measure-valued processes that we refer to as residual deadline related processes (see Lemma 5.3). Since the residual deadline related processes don't include information about the service discipline, the result in Lemma 5.3 may be of more general interest for analyzing queues with abandonment.

1.1. *Notation.* The following notation will be used throughout the paper. Let $\mathbb{N}$ denote the set of strictly positive integers and let $\mathbb{R}$ denote the set of real numbers. For $x, y \in \mathbb{R}$, $x \vee y = \max(x, y)$, $x \wedge y = \min(x, y)$, and $x^+ = x \vee 0$. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer less than or equal to x and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. For $x \in \mathbb{R}$, $\|x\| = |x|$ and for $x \in \mathbb{R}^2$, $\|x\| = \sqrt{x_1^2 + x_2^2}$.

The nonnegative real numbers $[0, \infty)$ will be denoted by $\mathbb{R}_+$. Let $\mathcal{B}_1$ denote the Borel subsets of $\mathbb{R}_+$ and $\mathcal{B}_2$ denote the Borel subsets of $\mathbb{R}_+^2$. On occasion we will use the notation $\mathbb{R}_+^1$ in place of $\mathbb{R}_+$. For a Borel set $B \in \mathcal{B}_i$, $i = 1, 2$, we denote the indicator function of the set B by 1_B. For $i = 1, 2$, $B \in \mathcal{B}_i$ and $x \in \mathbb{R}_+^i$, let B_x denote the x-shift of the set B, which is given by

$$B_x = \{y \in \mathbb{R}_+^i : y - x \in B\}. \tag{1}$$

Given $x \in \mathbb{R}_+$ and $B \in \mathcal{B}_2$, we adopt the following shorthand notation:

$$B_x = B_{(x,x)}.$$

For $i = 1, 2$, $B \in \mathcal{B}_i$, and $\kappa > 0$, let B^κ denote the κ-enlargement of B, which is given by

$$B^\kappa = \left\{ x \in \mathbb{R}_+^i : \inf_{y \in B} \|x - y\| < \kappa \right\}. \tag{2}$$

Notice that given $i = 1, 2$ and $B \in \mathcal{B}_i$, $(B^\kappa)_x \subseteq (B_x)^\kappa$, but $(B^\kappa)_x$ and $(B_x)^\kappa$ are not necessarily the same set. In particular, the set resulting from shifting before enlarging may contain additional points. We adopt the convention that $B_x^\kappa = (B_x)^\kappa$, i.e., B_x^κ is the larger of the two sets.

For $i = 1, 2$, let $\mathbf{C}_b(\mathbb{R}_+^i)$ denote the set of bounded continuous functions from $\mathbb{R}_+^i$ to $\mathbb{R}$. Given a finite, nonnegative Borel measure ζ on $\mathbb{R}_+^i$, let $\mathbf{L}(\zeta)$ denote the set of Borel measurable functions from $\mathbb{R}_+^i$ to $\mathbb{R}$ that are integrable with respect to ζ. For $g \in \mathbf{L}(\zeta)$, we define $\langle g, \zeta \rangle = \int_{\mathbb{R}_+^i} g \mathrm{d}\zeta$ and adopt the shorthand notation $\zeta(B) = \langle 1_B, \zeta \rangle$ for $B \in \mathcal{B}_i$. In addition, let $\chi : \mathbb{R}_+ \to \mathbb{R}$ be the identify map $\chi(x) = x$ for all $x \in \mathbb{R}_+$. Given a finite, nonnegative Borel measure ζ on $\mathbb{R}_+$, if $\chi \in \mathbf{L}(\zeta)$, i.e., if $\langle \chi, \zeta \rangle < \infty$, we say that ζ has a finite first moment.

For $i = 1, 2$, let $\mathbf{M}_i$ denote the set of finite, nonnegative Borel measures on $\mathbb{R}_+^i$. The zero measure in $\mathbf{M}_i$ is denoted by $\mathbf{0}$. For $x \in \mathbb{R}_+$, $\delta_x \in \mathbf{M}_1$

is the measure that puts one unit of mass at x. Similarly, for $x, y \in \mathbb{R}_+$, $\delta_{(x,y)} \in \mathbf{M}_2$ is the measure that puts one unit of mass at (x, y). The set $\mathbf{M}_i$ is endowed with the weak topology, that is, for $\zeta^n, \zeta \in \mathbf{M}_i$, $n \in \mathbb{N}$, we have $\zeta^n \xrightarrow{w} \zeta$ as $n \to \infty$ if and only if $\langle g, \zeta^n \rangle \to \langle g, \zeta \rangle$ as $n \to \infty$, for all $g \in \mathbf{C}_b(\mathbb{R}^i_+)$. With this topology, $\mathbf{M}_i$ is a Polish space [18]. Also define

$$\mathbf{M}_i^K = \{(\zeta_1, \ldots, \zeta_K) : \zeta_k \in \mathbf{M}_i \text{ for } 1 \le k \le K\}.$$

Then $\mathbf{M}_i^K$ endowed with the product topology is also a Polish space. Given $\zeta \in \mathbf{M}_i^K$ and $g \in \cap_{k=1}^K \mathbf{L}(\zeta_k)$, we define the shorthand notation

$$\langle g, \zeta \rangle = (\langle g, \zeta_1 \rangle, \ldots, \langle g, \zeta_K \rangle).$$

Given $i = 1, 2$ and $\zeta \in \mathbf{M}_i^K$, it will be handy to introduce the notation $\zeta_+ \in \mathbf{M}_i$ for the superposition measure, which is given by

$$\zeta_+(B) = \sum_{k=1}^K \zeta_k(B), \qquad \text{for all } B \in \mathcal{B}_i. \tag{3}$$

Let $\mathbf{M}_{1,0}$ denote the subset of $\mathbf{M}_1$ containing those measures that assign zero measure to the set $\{0\}$. Similarly, let $\mathbf{M}_{2,0}$ denote the subset of $\mathbf{M}_2$ containing those measures that assign measure zero to the set

$$C = \mathbb{R}_+ \times \{0\} \cup \{0\} \times \mathbb{R}_+. \tag{4}$$

For $x, y \in \mathbb{R}_+$, let $\delta_x^+ \in \mathbf{M}_{1,0}$ and $\delta_{(x,y)}^+ \in \mathbf{M}_{2,0}$ respectively denote the measures that put a unit mass at the point x if $x > 0$ and (x, y) if $x, y > 0$, and are the zero measure otherwise. For $i = 1, 2$, the set $\mathbf{M}_{i,0}^K$ is defined analogously to $\mathbf{M}_i^K$ except that $\zeta_k \in \mathbf{M}_{i,0}$ for $1 \le k \le K$. In addition, let $\mathcal{B}_{1,0}$ denote those sets in $\mathcal{B}_1$ that do not contain zero. Similarly, let $\mathcal{B}_{2,0}$ denote those sets in $\mathcal{B}_2$ that do not meet the set C.

Finally, we will use "$\Rightarrow$" to denote convergence in distribution of random elements of a metric space. Following Billingsley [4], we will use $\mathbb{P}$ and $\mathbb{E}$ respectively to denote the probability measure and expectation operator associated with whatever space the relevant random element is defined on. All stochastic processes used in this paper will be assumed to have paths that are right continuous with finite left limits (r.c.l.l.). For a Polish space $\mathcal{S}$, we denote by $\mathbf{D}([0, \infty), \mathcal{S})$ the space of r.c.l.l. functions from $[0, \infty)$ into $\mathcal{S}$, and we endow this space with the usual Skorohod J_1-topology (cf. [5]). There are six Polish spaces that will be considered in this paper: $\mathbb{R}$, $\mathbb{R}_+$, $\mathbf{M}_1$, $\mathbf{M}_1^K$, $\mathbf{M}_2$, and $\mathbf{M}_2^K$.

2. The Stochastic and Fluid Models.

2.1. *The Stochastic Model.* In this section, we define the model of the GI/GI/1 + GI queue serving K distinct customer classes, which will be used for the remainder of the paper.

Initial condition and associated system dynamics. The initial condition specifies the number $Z_+(0)$ of jobs in the queue at time zero, as well as the initial virtual sojourn time, initial patience time and class of each job. Assume that $Z_+(0)$ is nonnegative integer-valued random variable that is finite almost surely. The initial virtual sojourn times, initial patience times and classes are the first, second, and third coordinates respectively of the first $Z_+(0)$ elements of the random sequence $\{(\tilde{w}_j, \tilde{p}_j, k_j)\}_{j\in\mathbb{N}} \subset \mathbb{R}_+ \times \mathbb{R}_+ \times \{1, \ldots, K\}$. For $1 \le j \le Z_+(0)$, the initial job with initial virtual sojourn time $\tilde{w}_j$, initial patience time $\tilde{p}_j$, and class k_j is called job j.

We assume that the elements of the sequence $\{\tilde{w}_j\}_{j\in\mathbb{N}}$ are finite, positive, and nondecreasing, which reflects the fact that the service discipline is first-in-first-out. In particular, the job with the smallest index is regarded as having arrived to the system before all other jobs, and therefore has the smallest initial virtual sojourn time. This is the job currently in service and $\tilde{w}_1$ represents the time until its service is completed. The remaining jobs are waiting in the queue in order. For $2 \le j \le Z_+(0)$, $\tilde{w}_j$ represents the amount of time that job j will remain in the system, provided that it doesn't abandon before entering service.

We assume that the elements of the sequence $\{\tilde{p}_j\}_{j\in\mathbb{N}}$ are finite and positive, which reflects the fact that none of the jobs would be regarded as having abandoned the queue by time zero. For $1 \le j \le Z_+(0)$ such that $\tilde{p}_j > \tilde{w}_j$, job j is sufficiently patient to wait in the queue until service completion. For $1 \le j \le Z_+(0)$ such that $\tilde{p}_j \le \tilde{w}_j$, job j abandons the queue at time $\tilde{p}_j$ before entering service. We assume that $\tilde{w}_1 < \tilde{p}_1$, which reflects the fact that job 1 is presently in service and is therefore patient enough to stay in the queue until service completion. For $2 \le j \le Z_+(0)$, we assume that $\tilde{p}_j \le \tilde{w}_j$ if and only if $\tilde{w}_j = \tilde{w}_{j-1}$. When j is such that $\tilde{p}_j > \tilde{w}_j$, one regards $\tilde{w}_j - \tilde{w}_{j-1}$ as the service time of job j and this service time is included in the initial virtual sojourn time of all jobs ℓ such that $j \le \ell \le Z_+(0)$. When j is such that $\tilde{p}_j \le \tilde{w}_j$, a service time for job j is not included in any of the initial virtual sojourn times.

Since the queue is nonidling and first-in-first-out, all $Z_+(0)$ jobs in the system at time zero will either abandon or be served by time $W(0) = \tilde{w}_{Z_+(0)}$. We refer to $W(0)$ as the initial workload.

A convenient way to express the initial condition is to define an initial random measure $\mathcal{Z}(0) \in \mathbf{M}_2^K$. For this, we will find it convenient to separate the sequence $\{(\tilde{w}_j, \tilde{p}_j, k_j)\}_{j=1}^{Z_+(0)}$ into K separate sequences $\{(\tilde{w}_{k,j}, \tilde{p}_{k,j})\}_{j=1}^{Z_k(0)}$, one for each class. For $1 \le k \le K$, let $Z_k(0)$ denote the number of class k initial jobs. Given $1 \le k \le K$, for $1 \le j \le Z_k(0)$, let $i(k,j)$ be the jth smallest index such that $k_{i(k,j)} = k$ and set $\tilde{w}_{k,j} = \tilde{w}_{i(k,j)}$ and $\tilde{p}_{k,j} = \tilde{p}_{i(k,j)}$. Then, for $1 \le k \le K$, let $\mathcal{Z}_k(0) \in \mathbf{M}_2$ be given by

$$\mathcal{Z}_k(0) = \sum_{j=1}^{Z_k(0)} \delta^+_{(\tilde{w}_{k,j}, \tilde{p}_{k,j})},$$

which equals $\mathbf{0}$ if $Z_k(0) = 0$. Then let

$$\mathcal{Z}(0) = (\mathcal{Z}_1(0), \dots, \mathcal{Z}_K(0)).$$

Our assumptions imply that $\mathcal{Z}(0)$ satisfies

$$\mathbf{P}\left(\max_{1\le k\le K} \langle 1, \mathcal{Z}_k(0)\rangle \vee W(0) < \infty\right) = 1. \tag{5}$$

Furthermore,

$$\mathbf{P}\left(\max_{1\le k\le K} \mathcal{Z}_k(0)(C) = 0\right) = 1.$$

In particular, $\mathcal{Z}(0) \in \mathbf{M}_{2,0}^K$ almost surely.

Stochastic primitives and associated system dynamics. The stochastic primitives consist of K exogenous arrival processes $E_k(\cdot)$, $1 \le k \le K$, K sequences of service times $\{v_{k,i}\}_{i\in\mathbb{N}}$, $1 \le k \le K$, and K sequences of deadlines $\{d_{k,i}\}_{i\in\mathbb{N}}$, $1 \le k \le K$. We assume that the exogenous arrival processes, the sequences of service times, and the sequences of deadlines are all independent of one another.

For a given $1 \le k \le K$, the class k arrival process $E_k(\cdot)$ is a rate $\lambda_k \in (0,\infty)$ renewal process. For $t \in [0,\infty)$, $E_k(t)$ represents the number of class k jobs that arrive to the queue during the time interval $(0,t]$. We assume that the interarrival times are strictly positive and denote the sequence of interarrival times by $\{\xi_{k,i}\}_{i\in\mathbb{N}}$. Class k jobs arriving after time zero are indexed by integers $j > Z_k(0)$. For $t \in [0,\infty)$, let

$$A_k(t) = Z_k(0) + E_k(t). \tag{6}$$

Then class k job j arrives at time $t_{k,j} = \inf\{t \in [0,\infty) : A_k(t) \ge j\}$. Hence, for $j' < j$, $t_{k,j'} \le t_{k,j}$ and we say that class k job j' arrives before class

k job j. The inequality is strict for indices $j > Z_k(0)$. Moreover, for each $j \le Z_k(0)$, $t_{k,j} = 0$.

Given $1 \le k \le K$, for each $i \in \mathbb{N}$, the random variable $v_{k,i}$ represents the service time of the $(Z_k(0) + i)$th class k job. That is, class k job $j > Z_k(0)$ has service time $v_{k,j-Z_k(0)}$. Assume that the random variables $\{v_{k,i}\}_{i\in\mathbb{N}}$ are strictly positive and form an independent and identically distributed sequence with finite positive mean $1/\mu_k$. Define the class k offered load to be $\rho_k = \lambda_k/\mu_k$. We assume that the queue is overloaded. In particular, we assume that $\rho = \sum_{k=1}^K \rho_k > 1$.

Given $1 \le k \le K$, for each $i \in \mathbb{N}$, the random variable $d_{k,i}$ represents the deadline of the $(Z_k(0) + i)$th class k job. That is, class k job $j > Z_k(0)$ has deadline $d_{k,j-Z_k(0)}$. Assume that the random variables $\{d_{k,i}\}_{i\in\mathbb{N}}$ are strictly positive and form an independent and identically distributed sequence of random variables with common continuous distribution Γ_k. Assume that the mean $1/\gamma_k$ is finite. Let $F_k(\cdot)$ denote the cumulative distribution function associated with Γ_k, i.e., $F_k(x) = \langle 1_{[0,x]}, \Gamma_k\rangle$ for all $x \in \mathbb{R}_+$. Denote its complement by $G_k(\cdot) = 1 - F_k(\cdot)$.

As is the case for jobs in the system at time zero, jobs arriving after time zero are served in a first-in-first-out, nonidling fashion. A class k job j arriving to the system after time zero immediately enters service if the server is available. Otherwise, for class k job j to be served, it must wait until all other jobs currently in the queue exit via service completion or abandonment. If class k job j has not entered service before time $t_{k,j} + d_{k,j-Z_k(0)}$, class k job j abandons the queue at this time. Otherwise, when class k job j enters service, it is served for $v_{k,j-Z_k(0)}$ time units.

It will be convenient to combine the exogenous arrival process and deadlines into a single measure-valued deadline process.

DEFINITION 2.1. For $1 \le k \le K$, the class k deadline process is given by

$$\mathcal{D}_k(t) = \sum_{i=1}^{E_k(t)} \delta_{d_{k,i}}, \quad t \in [0,\infty).$$

Then the deadline process is given by

$$\mathcal{D}(t) = (\mathcal{D}_1(t), \ldots, \mathcal{D}_K(t)), \quad t \in [0,\infty).$$

Note that $\mathcal{D}_k(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M}_1)$ for $1 \le k \le K$ and $\mathcal{D}(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M}_1^K)$.

The workload process. The workload process $W(\cdot) \in \mathbf{D}([0,\infty), \mathbb{R}_+)$ tracks as a function of time, the amount of time needed for the server to process all jobs currently in the system that will not abandon. If no additional jobs were to arrive after time t, the system would be empty $W(t)$ time units in the future. This quantity also records the amount of time that a newly arriving job would have to wait before being served, if the job were sufficiently patient. In particular, as in [11], $W(\cdot)$ almost surely satisfies, for all $t \in [0,\infty)$,

$$(7)\quad W(t) = W(0) + \sum_{k=1}^{K} \int_{(0,t]} v_{k,E_k(s)} 1_{\{d_{k,E_k(s)} > W(s-)\}} \mathrm{d}E_k(s) - B(t),$$

$$(8)\quad B(t) = \int_0^t 1_{\{W(s)>0\}} \mathrm{d}s.$$

Here $B(\cdot)$ denotes the busy time process. Since the server works at rate one during any busy period, the amount of work served by time t is equal to $B(t)$. Occasionally, it will be convenient to refer to the idle time process, which is given by $I(t) = t - B(t)$ for $t \in [0,\infty)$. For a given time t, the first and second terms in (7) add up all of the work that enters the system by time t and doesn't abandon before entering service. Note that because of the indicator in the integrand, a particular job's service time is added to the workload if and only if that job will not abandon before it enters service. Fluid and diffusion limit results for this process were proved in [11], where it was referred to as the virtual waiting time process. We will leverage that fluid limit result to carry out the analysis here.

Evolution of the virtual sojourn times and patience times. For $1 \le k \le K$, let

$$w_{k,j} = \begin{cases} \tilde{w}_{k,j}, & 1 \le j \le Z_k(0), \\ W(t_{k,j}), & j > Z_k(0), \end{cases}$$

$$p_{k,j} = \begin{cases} \tilde{p}_{k,j}, & 1 \le j \le Z_k(0), \\ d_{k,j-Z_k(0)} + v_{k,j-Z_k(0)} 1_{\{d_{k,j-Z_k(0)} > W(t_{k,j}-)\}}, & j > Z_k(0). \end{cases}$$

Note that for each $1 \le k \le K$ and $j > Z_k(0)$,

$$W(t_{k,j}) = W(t_{k,j}-) + v_{k,j-Z_k(0)} 1_{\{d_{k,j-Z_k(0)} > W(t_{k,j}-)\}}.$$

Hence, if class k job j is served, $w_{k,j}$ includes the job's service time in addition to the time spent waiting for service to begin. So, if class k job j is served, it stays in the system $w_{k,j}$ time units. Therefore, $w_{k,j}$ is referred to as the virtual sojourn time of class k job j. Further, $p_{k,j}$ represents the initial patience of class k job j. If a class k job j arriving after time zero will enter service before time $t_{k,j} + d_{k,j-Z_k(0)}$, the job's patience time is taken to be $d_{k,j-Z_k(0)} + v_{k,j-Z_k(0)}$ to account for the fact that it will not abandon $d_{k,j-Z_k(0)}$ time units after arrival. Instead, it will stay until time $t_{k,j} + w_{k,j}$ when it receives its full service time requirement. Otherwise, if class k job j won't enter service before the deadline expires, the job will abandon at time $t_{k,j} + d_{k,j-Z_k(0)}$, and $p_{k,j} = d_{k,j-Z_k(0)}$. Then for each $1 \le k \le K$ and $j \in \mathbb{N}$, the sojourn time $s_{k,j}$ of class k job j is given by

$$s_{k,j} = \min(w_{k,j}, p_{k,j}).$$

This quantity indicates precisely how long class k job j will reside in the system.

For all $t \in [0,\infty)$, $1 \le k \le K$, and $1 \le j \le A_k(t)$, define

$$w_{k,j}(t) = (w_{k,j} - (t - t_{k,j}))^+, \tag{9}$$

$$p_{k,j}(t) = (p_{k,j} - (t - t_{k,j}))^+. \tag{10}$$

For $1 \le k \le K$, $j \in \mathbb{N}$ and $t \ge t_{k,j}$, $w_{k,j}(t)$ and $p_{k,j}(t)$ respectively represent the residual virtual sojourn time and residual patience time of class k job j at time t. Then the residual sojourn time $s_{k,j}(t)$ for class k job j at time $t \ge t_{k,j}$ is given by

$$s_{k,j}(t) = \min(w_{k,j}(t), p_{k,j}(t)).$$

Measure-valued state descriptor. For $1 \le k \le K$, define the class k state descriptor by

$$\mathcal{Z}_k(t) = \sum_{j=1}^{A_k(t)} \delta^+_{(w_{k,j}(t),p_{k,j}(t))}, \qquad t \in [0,\infty). \tag{11}$$

The state descriptor is defined as

$$\mathcal{Z}(t) = (\mathcal{Z}_1(t), \ldots, \mathcal{Z}_K(t)), \qquad t \in [0,\infty). \tag{12}$$

For each $1 \le k \le K$, $\mathcal{Z}_k(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M}_2)$, and $\mathcal{Z}(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M}_2^K)$.

Figure 1 depicts one component of a hypothetical system state at a fixed time. The points in the figure correspond to unit atoms of the measure. All points move to the left and down at rate one. The dotted diagonal line $p = w$ separates the points in the figure into two groups. The jobs associated with points on or below the line will eventually abandon; the points above the line represent jobs that will be served. Once a point reaches one of the coordinate axes it immediately leaves the system and so is no longer included in the system state. Among all of the components of the system state, there is a

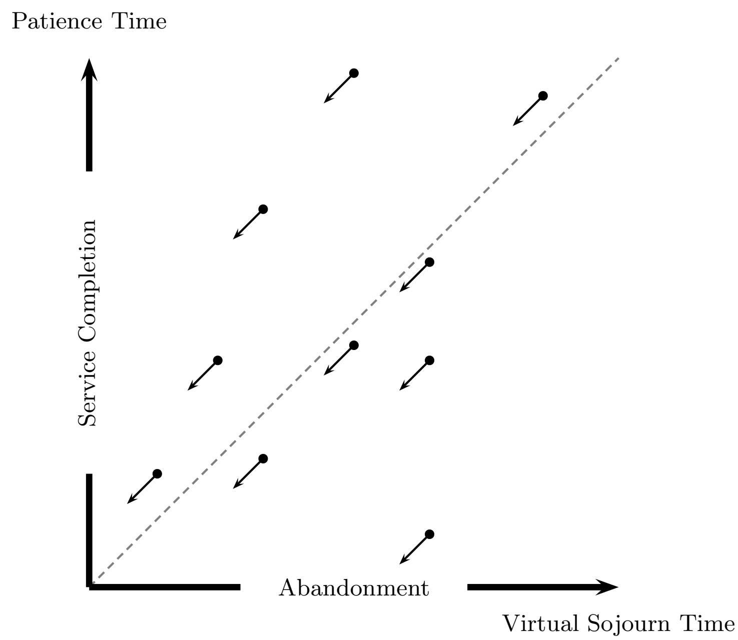

FIG 1. *One coordinate of a hypothetical system state at a fixed time with depicted transitions.*

unique point that has the smallest residual virtual sojourn time coordinate and lies above the diagonal. This point corresponds to the job currently in service.

Notice that in Figure 1 there are three sets of points that are aligned vertically. In each set, at most one of these points is above the dotted line and the corresponding job will be served. For each residual virtual sojourn time assumed by some job that is in the system at this fixed time, there is exactly one of these jobs for which the location of the corresponding point in the component of the state descriptor associated with that job's class lies above the diagonal. This job will be served and it arrived before any other job with the same residual virtual sojourn time. These later arriving jobs aren't patient enough to remain in queue until service begins. Instead each will abandon. Therefore the corresponding points in the components of the state descriptor associated with those jobs' classes lie on or below the diagonal. Although their sequence of arrivals relative to one another is not captured by the state descriptor, the relative residual patience times reveal the order in which they will abandon.

The state descriptor satisfies the following system of dynamic equations. For each $1 \le k \le K$ and for all $B \in \mathcal{B}_{2,0}$ and $t \in [0,\infty)$,

$$\mathcal{Z}_k(t)(B) = \sum_{j=1}^{A_k(t)} 1_{B_{t-t_{k,j}}}(w_{k,j}, p_{k,j}). \tag{13}$$

To see this, simply note that for each $B \in \mathcal{B}_{2,0}$, $x \in \mathbb{R}_+$, and $w, p > 0$,

$$((w-x)^+, (p-x)^+) \in B \quad \Leftrightarrow \quad (w-x, p-x) \in B \quad \Leftrightarrow \quad (w,p) \in B_x.$$

The dynamic equation (13) is equivalent to

$$\mathcal{Z}_k(t)(B) = \mathcal{Z}_k(0)(B_t) + \sum_{j=1+Z_k(0)}^{A_k(t)} 1_{B_{t-t_{k,j}}}(w_{k,j}, p_{k,j}). \tag{14}$$

Given $B \in \mathcal{B}_{2,0}$ and $x, y \in \mathbb{R}_+$, notice that

$$(B_x)_y = B_{x+y}.$$

This together with (13) implies that, for each $1 \le k \le K$ and for all $B \in \mathcal{B}_{2,0}$ and $h, t \in [0,\infty)$,

$$\mathcal{Z}_k(t+h)(B) = \mathcal{Z}_k(t)(B_h) + \sum_{j=A_k(t)+1}^{A_k(t+h)} 1_{B_{t+h-t_{k,j}}}(w_{k,j}, p_{k,j}). \tag{15}$$

2.2. *The Fluid Model.* In this section, we define the fluid model associated with this GI/GI/1+GI queue with K distinct job classes. The primitive data for this fluid model consists of the vector λ of arrival rates, the vector μ of service rates, and the vector Γ of deadline distributions. The triple (λ, μ, Γ) is referred to as supercritical data since $\rho > 1$. We begin by summarizing the results in [11] that suggest a workload fluid model. Then, we develop a full measure-valued fluid model. For this, we first need to define the scaling regimes that yield the desired limiting dynamics.

The Sequence of Time Accelerated Systems. We consider a sequence of systems indexed by $n \in \mathbb{N}$ in which the arrival rates and mean service times in the nth system are sped up by a factor of n. We use a superscript n to denote all processes and parameters associated with the nth system.

Specifically, the interarrival times $\{\xi^n_{k,i}\}_{i\in\mathbb{N}}$ for class k in the nth system are given by $\xi^n_{k,i} = \xi_{k,i}/n$ for $i \in \mathbb{N}$. Then $E^n_k(\cdot)$ denotes the class k exogenous arrival processes in the nth system. Hence, for $1 \le k \le K$, $E^n_k(t) = E_k(nt)$

for $t \in [0,\infty)$ and $\lambda_k^n = n\lambda_k$. Similarly, the class k service times $\{v_{k,i}^n\}_{i\in\mathbb{N}}$ in the nth system are given by $v_{k,i}^n = v_{k,i}/n$ for $i \in \mathbb{N}$. Then $\mu_k^n = n\mu_k$ and $\rho_k^n = \rho_k$. Hence, $\rho^n = \rho$.

The deadline sequence is unscaled. In particular, for each $1 \le k \le K$ and $n, i \in \mathbb{N}$, $d_{k,i}^n = d_{k,i}$. Hence, we omit the superscript when referring to the class k deadlines, distribution Γ_k, mean $1/\gamma_k$, and cumulative distribution function F_k or its complement G_k. Then the class k deadline process for the nth system is given by

$$\mathcal{D}_k^n(t) = \sum_{i=1}^{E_k^n(t)} \delta_{d_{k,i}}, \qquad t \in [0,\infty),$$

and $\mathcal{D}^n(\cdot) = (\mathcal{D}_1^n(\cdot), \ldots, \mathcal{D}_K^n(\cdot))$.

For each $n \in \mathbb{N}$, there is an initial condition $Z^n(0)$ and $\{(\tilde{w}_j^n, \tilde{p}_j^n, k_j^n)\}_{j\in\mathbb{N}}$ satisfying the conditions indicated above. The initial conditions may vary with n. Then for $n \in \mathbb{N}$, $1 \le k \le K$, and $t \in [0,\infty)$, $A_k^n(t) = Z_k^n(0) + E_k^n(t)$. If we suppose for simplicity that $Z_+(0) = 0$ in the unscaled system, then the job arrival times $\{t_{k,j}^n\}_{j\in\mathbb{N}}$ for class k jobs in the nth system would be given by

$$t_{k,j}^n = \begin{cases} 0, & 1 \le j \le Z_k^n(0), \\ \frac{t_{k,j-Z_k^n(0)}}{n}, & j > Z_k^n(0). \end{cases}$$

The workload $W^n(\cdot)$, busy time $B^n(\cdot)$, and idle time $I^n(\cdot)$ processes for the nth system satisfy equations analogous to (7) and (8). The residual virtual sojourn $w_{k,j}^n(\cdot)$ and residual patience $p_{k,j}^n(\cdot)$ times, $1 \le k \le K$ and $j \in \mathbb{N}$, for the nth system are defined as in (9) and (10). The class k state descriptor $\mathcal{Z}_k^n(\cdot)$ and the state descriptor $\mathcal{Z}^n(\cdot)$ for the nth system are defined as in (11) and (12). Analogs of (13) and (15) hold for the nth system as well.

The Workload Fluid Model. For the time accelerated scaling regime, the authors of [11] identify what one might refer to as a workload fluid model. Namely, define a *workload fluid model solution* to be a function $w : [0,\infty) \to \mathbb{R}_+$ satisfying

$$w(t) = w(0) + \sum_{k=1}^{K} \rho_k \int_0^t G_k(w(s))\mathrm{d}s - t, \qquad t \in [0,\infty). \tag{16}$$

Equation (16) can be interpreted as a fluid analog of (7) and (8). Notice that any solution to (16) is necessarily Lipschitz continuous with Lipschitz constant $\rho - 1$, and therefore is almost everywhere differentiable. In [11], it is asserted that for each $w_0 \in \mathbb{R}_+$ a unique workload fluid model solution

$w(\cdot)$ with $w(0) = w_0$ exists. They further show that workload fluid model solutions satisfy a nice monotonicity property [11, Theorem 1]. To describe this, let

$$\begin{aligned} w_l &= \sup\left\{u \in \mathbb{R}_+ : \sum_{k=1}^K \rho_k G_k(u) < 1\right\}, \\ w_u &= \sup\left\{u \in \mathbb{R}_+ : \sum_{k=1}^K \rho_k G_k(u) \leq 1\right\}. \end{aligned}$$

Note that by continuity of $G_k(\cdot)$ for each k and the fact that $\rho > 1$, $0 < w_l \leq w_u < \infty$. They show that for a workload fluid model solution $w(\cdot)$ such that $w(0) \notin [w_l, w_u]$, $w(\cdot)$ is strictly monotone and

$$\lim_{t\to\infty} w(t) = \begin{cases} w_l, & \text{if } w(0) < w_l, \\ w_u, & \text{if } w(0) > w_u. \end{cases} \tag{17}$$

Otherwise, if $w(0) \in [w_l, w_u]$, then $w(t) = w(0)$ for all $t \in [0, \infty)$.

Workload fluid model solutions also satisfy a useful relative ordering property. In particular, given two workload fluid model solutions $w_1(\cdot)$ and $w_2(\cdot)$ such that $w_1(0) \leq w_2(0)$ it follows that for all $t \in [0, \infty)$,

$$w_1(t) \leq w_2(t). \tag{18}$$

To see this note that if $w_1(0) \leq w_u$ and $w_l \leq w_2(0)$, then (18) follows immediately from the monotoncity property of workload fluid model solutions. Otherwise, $w_2(0) < w_l$ or $w_1(0) > w_u$. If $w_1(0) > w_u$, then by continuity and (17) there exists $t \in [0, \infty)$ such that $w_2(t) = w_1(0)$. For $s \in [0, \infty)$, set $w(s) = w_2(t+s)$. Then, it is straightforward to verify that $w(\cdot)$ is a workload fluid model solution with $w(0) = w_1(0)$. Hence, by uniqueness of fluid model solutions, $w(\cdot) = w_1(\cdot)$ so that $w_2(\cdot + t) = w_1(\cdot)$. Then for all $s \in [0, \infty)$, $w_1(s) = w_2(s+t) \leq w_2(s)$. An analogous argument demonstrates (18) when $w_2(0) < w_l$.

In [11], the authors justify interpreting the fluid model as a first order approximation of the stochastic system by showing that if $W^n(0) \to w_0$ almost surely as $n \to \infty$, where w_0 is a finite nonnegative constant, then $(W^n, I^n) \Rightarrow (w(\cdot), 0)$, as $n \to \infty$, where $w(\cdot)$ is the unique workload fluid model solution such that $w(0) = w_0$ [11, Theorem 1]. Here, we wish to cite a slightly modified version of [11, Theorem 1], which we state next. Before this, we make note of a representation for the workload process developed in [11] that will be of use here. By [11, (19) and arguments presented in the proof of Theorem 1], for all $n \in \mathbb{N}$ and $0 \leq s \leq t < \infty$,

$$
\begin{aligned}
W^n(t) &= W^n(s) + I^n(t) - I^n(s) - (t-s) \\
&\quad + X^n(t) - X^n(s) + \sum_{k=1}^{K} \rho_k \int_s^t G_k(W^n(u))\mathrm{d}u,
\end{aligned} \tag{19}
$$

where $\{X^n(\cdot)\}_{n\in\mathbb{N}} \subset \mathbf{D}([0,\infty),\mathbb{R})$ is such that as $n \to \infty$,

$$X^n(\cdot) \Rightarrow 0. \tag{20}$$

THEOREM 2.2. *If $W^n(0) \Rightarrow W_0$ as $n \to \infty$, then as $n \to \infty$,*

$$(W^n, I^n) \Rightarrow (W^*(\cdot), 0), \tag{21}$$

where $W^*(0)$ is equal in distribution to W_0. Furthermore, $W^*(\cdot)$ is almost surely a workload fluid model solution.

SUMMARY OF PROOF. Let $\iota : [0,\infty) \to [0,\infty)$ be given by $\iota(t) = t$ for all $t \in [0,\infty)$. Since the limit in (20) is deterministic, it follows that as $n \to \infty$,

$$W^n(0) - \iota(\cdot) + X^n(\cdot) \Rightarrow W_0 - \iota(\cdot).$$

Using tools developed in [19], the authors further show that for all $n \in \mathbb{N}$,

$$(W^n(\cdot), I^n(\cdot)) = (\varphi_G, \psi_G)\left(W^n(0) - \iota(\cdot) + X^n(\cdot)\right), \tag{22}$$

where $(\varphi_G, \psi_G) : \mathbf{D}([0,\infty),\mathbb{R}) \to \mathbf{D}([0,\infty),\mathbb{R}_+^2)$ is a Lipschitz continuous function in the topology of uniform convergence on compact sets. Then, by the continuous mapping theorem, it follows that as $n \to \infty$,

$$(W^n(\cdot), I^n(\cdot)) \Rightarrow (\varphi_G, \psi_G)\left(W_0 - \iota(\cdot)\right). \tag{23}$$

The authors of [11] go on to show that for $w_0 \in \mathbb{R}_+$ $(\varphi_G, \psi_G)(w_0 - \iota(\cdot)) = (w(\cdot), 0)$, where $w(\cdot)$ is the unique workload fluid model solution such that $w(0) = w_0$. □

The Measure-Valued Fluid Model. To develop a measure-valued fluid model, we add spatial scaling to the sequence of time accelerated systems. In particular in order to obtain a fluid scaled system, we divide space in the nth time accelerated system by a factor of n. Then, for $n \in \mathbb{N}$, $1 \leq k \leq K$, and $t \in [0,\infty)$,

$$\bar{E}_k^n(t) = \frac{E_k^n(t)}{n}, \qquad \bar{\mathcal{D}}_k^n(t) = \frac{\mathcal{D}_k^n(t)}{n}, \qquad \text{and} \qquad \bar{\mathcal{Z}}_k^n(t) = \frac{\mathcal{Z}_k^n(t)}{n}.$$

The nth fluid scaled state descriptor in the sequence is given by

$$\bar{\mathcal{Z}}^n(\cdot) = (\bar{\mathcal{Z}}_1^n(\cdot), \ldots, \bar{\mathcal{Z}}_K^n(\cdot)).$$

We are interested the the limiting behavior as n approaches infinity.

One can speculate as to how the dynamics of the system will behave in the limit. Class k fluid should arrive to the system at rate λ_k and be distributed over the positive quadrant as it enters. Class k fluid arriving at time t should have virtual sojourn time $w(t)$, where $w(\cdot)$ is an appropriate fluid workload solution, and patience time distributed according to Γ_k. Then the residual virtual sojourn and residual patience times should each decrease at rate one until at least one is zero, at which time the fluid would exit from the system. In particular, at time $t + h$, fluid that arrived to (w, p) at time t should be at $(w-h, p-h)$ if $(w-h) \wedge (p-h) > 0$ and should have exited the system otherwise. Such departures would be associated with abandonment if $p - h \leq 0 \wedge (w-h)$ and service completion if $w - h \leq 0$ and $w - h < p - h$. We wish to capture these dynamics by defining an appropriate fluid model.

The initial measure will need to satisfy various natural conditions. To define these, first recall the definition of C given in (4). Then, given $\vartheta \in \mathbf{M}_2^K$, recall that $\vartheta_+ \in \mathbf{M}_2$ denotes the superposition measure (see (3)). Let

$$w_\vartheta = \sup\{x \in \mathbb{R}_+ : \vartheta_+([x, \infty) \times \mathbb{R}_+) > 0\}.$$

Define $\mathbf{I} \subset \mathbf{M}_2^K$ to be the collection of $\vartheta \in \mathbf{M}_2^K$ such that

(I.1) $\vartheta_+(C_x) = 0$ for all $x \in \mathbb{R}_+^2$,
(I.2) $w_\vartheta < \infty$,
(I.3) $\max_{1 \leq k \leq K} G_k(w_\vartheta - \varepsilon) > 0$ for all $\varepsilon > 0$.

We refer to the collection $\mathcal{C} = \{C_x : x \in \mathbb{R}_+^2\}$ as the corner sets. Then condition (I.1) is that each coordinate of the initial measure doesn't charge corner sets, which is equivalent to the condition that each coordinate of the initial measure doesn't charge vertical or horizontal lines. A motivation for requiring (I.1) is to prevent exiting mass from resulting in discontinuities. Since (I.1) it is preserved by the fluid model dynamics (see Property 2 of Theorem 3.1), it is a natural to require it of the initial condition.

Given $\vartheta \in \mathbf{M}_2^K$ that satisfies (I.1) and $\varepsilon > 0$ there exists $\kappa > 0$ such that

$$\max_{1 \leq k \leq K} \sup_{x \in \mathbb{R}_+^2} \vartheta_k(C_x^\kappa) < \varepsilon. \tag{24}$$

To see this, given $x \in \mathbb{R}_+^2$, let $p_i(x) = x_i$ for $i = 1, 2$. Then, given $\nu \in \mathbf{M}_2$ and $i = 1, 2$, let $\pi_i(\nu) \in \mathbf{M}_1$ be the projection measure such that for all

$f \in \mathbf{C}_b(\mathbb{R}^1_+)$

$$\langle f, \pi_i(\nu)\rangle = \langle f \circ p_i, \nu\rangle \,. \tag{25}$$

Then, if $\vartheta \in \mathbf{M}_2^K$ satisfies (I.1), it follows that for $i = 1, 2$, $\pi_i(\vartheta_+)$ doesn't charge points. Hence, for every $\varepsilon > 0$ there exists $\kappa > 0$ such that

$$\max_{i=1,2} \sup_{y \in \mathbb{R}_+} \pi_i(\vartheta_+)([(y-\kappa)^+, y+\kappa]) < \frac{\varepsilon}{2},$$

(cf. [8, Lemma A.1]). Since for each $1 \le k \le K$ and $x \in \mathbb{R}^2_+$,

$$\vartheta_k(C_x^\kappa) \le \vartheta_+(C_x^\kappa) \le \pi_1(\vartheta_+)([(x_1-\kappa)^+, x_1+\kappa]) + \pi_2(\vartheta_+)([(x_2-\kappa)^+, x_2+\kappa]),$$

(24) follows.

Condition (I.2) dictates that the fluid analog of the initial workload is finite and condition (I.3) requires that the initial fluid workload does not exceed the maximal deadline. In particular, let

$$d_{\max} = \max_{1 \le k \le K} \sup\{x \in \mathbb{R}_+ : G_k(x) > 0\}.$$

Then, by (I.2), $w_\vartheta < d_{\max}$ if $d_{\max} = \infty$. Further, by (I.3), $w_\vartheta \le d_{\max}$ if $d_{\max} < \infty$, which is natural. Indeed in the stochastic system the workload can only exceed the maximal deadline by at most one service time, i.e., once the workload has jumped above the maximal deadline, all incoming jobs necessarily abandon until such time as the maximal deadline exceeds the workload. This discrepancy vanishes in the fluid limit. The reader will see that it is needed mathematically when $d_{\max} < \infty$ to prevent fluid from building up on a moving vertical line during $(0, w_\vartheta - d_{\max}]$, which would prevent (I.1) from being preserved.

A *fluid model solution* for the initial measure $\vartheta \in \mathbf{I}$ is a function $\zeta : [0,\infty) \to \mathbf{M}_2^K$ such that $\zeta(0) = \vartheta$ and for each $1 \le k \le K$, $B \in \mathcal{B}_2$, and $t \in [0,\infty)$, ζ_k satisfies

$$\zeta_k(t)(B) = \zeta_k(0)(B_t) + \lambda_k \int_0^t \left(\delta^+_{w(s)} \times \Gamma_k\right)(B_{t-s})\mathrm{d}s, \tag{26}$$

where $w(\cdot)$ denotes the unique workload fluid model solution with $w(0) = w_\vartheta$. Notice that (26) is a fluid analog of (13).

3. Main Results and Approximation Formulas. Here we state the two main theorems proved in this paper (Theorems 3.1 and 3.2), which together validate approximating various performance processes via fluid model solutions. Then, we derive some specific approximation formulas. Lastly, we identify the set of invariant states for the fluid model.

THEOREM 3.1. *Let $\vartheta \in \mathbf{I}$. Then there exists a unique fluid model solution $\zeta(\cdot)$ for the data (λ, μ, Γ) such that $\zeta(0) = \vartheta$. In addition,*

1. *$w_{\zeta(t)} = w(t)$ for all $t \in [0, \infty)$, where $w(\cdot)$ is the unique workload fluid model solution such that $w(0) = w_\vartheta$;*
2. *$\zeta_+(t)(C_x) = 0$ for all $x \in \mathbb{R}^2_+$ and $t \in [0, \infty)$;*
3. *$\zeta(\cdot)$ is continuous.*

The proof of this theorem is given in Section 4. Property 1 is that the right edge of the support of the fluid model solution superposition measure is equal to the workload fluid model solution for all time, which is true by definition at time zero. It implies that (I.2) holds for all time. Then by the monotoncity properties of workload fluid model solutions, (I.3) holds for all time as well. The proof of Property 1 is fairly straightforward, as the reader will see in Section 4. Property 2 is that the fluid model solution superposition measure doesn't charge corner sets for all time, i.e., that (I.1) holds for all time, and its proof is more involved. For this, we first prove Lemma 4.1, from which Property 2 follows by letting ε decrease to zero. Lemma 4.1 is then used together with some additional arguments to verify Property 3, continuity.

The next result justifies regarding fluid model solutions as first order approximations for the measure-valued state descriptor of the original stochastic system.

THEOREM 3.2. *Suppose that $\mathcal{Z}_0^* = (\mathcal{Z}_{0,1}^*, \ldots, \mathcal{Z}_{0,K}^*)$ is a random measure in $\mathbf{M}_2^K$ with superposition measure $\mathcal{Z}_{0,+}^*$ such that*

(A.1) $\mathbb{P}(\mathcal{Z}_0^* \in \mathbf{I}) = 1$,
(A.2) $\mathbb{E}[\langle p_1 + p_2, \mathcal{Z}_{0,+}^* \rangle] < \infty$,
(A.3) $\mathbb{E}[\mathcal{Z}_{0,+}^*(\mathbb{R}^2_+)] < \infty$.

Let $W_0^* = w_{\mathcal{Z}_0^*}$. Also suppose that, as $n \to \infty$,

$$\begin{aligned}(27)\qquad &\left(\bar{\mathcal{Z}}^n(0), \left\langle p_1, \bar{\mathcal{Z}}^n(0)\right\rangle, \left\langle p_2, \bar{\mathcal{Z}}^n(0)\right\rangle, W^n(0)\right)\\ &\quad \Rightarrow \left(\mathcal{Z}_0^*, \langle p_1, \mathcal{Z}_0^*\rangle, \langle p_2, \mathcal{Z}_0^*\rangle, W_0^*\right).\end{aligned}$$

Then, as $n \to \infty$,

$$\bar{\mathcal{Z}}^n(\cdot) \Rightarrow \mathcal{Z}^*(\cdot),$$

where $\mathcal{Z}^*(0)$ is equal in distribution to $\mathcal{Z}_0^*$. Furthermore, $\mathcal{Z}^*(\cdot)$ is almost surely a fluid model solution.

This result is proved via verifying tightness and then proving that any limit point is almost surely a fluid model solution, which are done in Sections 5 and 6 respectively.

Next, we use the fluid model to derive various approximation formulas. These include approximation formulas that demonstrate the nonlinear nature of state space collapse that takes place for this model. That is to say, one can recover various K dimensional quantities such as the fluid approximation for the queue-length vector from the fluid approximation for the workload process. However, this mapping depends on the entirety of the deadline distributions and is therefore nonlinear in its dependence on the workload process. See (30), (31), and (32) below. In addition, we are able to approximate various age related processes such as the number of jobs in the system that are within a specific age range.

To describe these, we will need to introduce the following functional. Given $\vartheta \in \mathbf{I}$, let $\zeta(\cdot)$ denote the unique fluid model solution such that $\zeta(0) = \vartheta$. Let $w(\cdot)$ denote the unique workload fluid model solution such that $w(0) = w_\vartheta$. Fluid that arrives at time $s > 0$ has fluid workload $w(s)$. Some of this fluid remains in the system at time $t \geq s$ only if

$$w(s) - (t - s) > 0.$$

Recall that $w(\cdot)$ is continuous. Furthermore, it is strictly decreasing if $w(0) > w_u$ and bounded above by w_u otherwise. Additionally, since $\rho > 1$, $w_u < d_{\max}$. Finally, by (I.3), $w(0) \leq d_{\max}$. Hence, $w(s) < d_{\max}$ for all $s > 0$. Therefore, $w'(s) = \sum_{k=1}^{K} \rho_k G_k(w(s)) - 1 > -1$ for all $s > 0$. Then, as a function of $s \in [0, \infty)$, $w(s) + s$ is continuous and strictly increasing. Furthermore, for $s \in [0, t]$, the values range from $w(0)$ to $w(t) + t \geq t$ at time t. Hence, $\inf\{0 \leq s \leq t : w(s) + s \geq t\}$ is well defined, finite, and can be interpreted as the time at which fluid departing at time t via service completion arrived to the system. For $t \in [0, \infty)$, let

$$\tau(t) = \inf\{0 \leq s \leq t : w(s) + s \geq t\}. \tag{28}$$

Then $\tau(t) = 0$ for all $t \in [0, w(0)]$, so that $w(\tau(t)) + \tau(t) > t$ for all $t \in [0, w(0))$. Further, for $t \geq w(0)$,

$$w(\tau(t)) + \tau(t) = t. \tag{29}$$

For $t > w(0)$, all fluid in the system at time t arrived during the time interval $(\tau(t), t]$.

Nonlinear State Space Collapse. Given $\vartheta \in \mathbf{I}$, let $\zeta(\cdot)$ denote the unique fluid model solution with $\zeta(0) = \vartheta$ and $w(\cdot)$ denote the unique workload fluid model solution with $w(0) = w_\vartheta$.

Queue-Length Vector Fluid Approximation. For $1 \le k \le K$ and $t \in [0, \infty)$,

$$z_k(t) \equiv \zeta_k(t)(\mathbb{R}_+^2) = \begin{cases} \zeta_k(0)((\mathbb{R}_+^2)_t) + \lambda_k \int_0^t G_k(a)\mathrm{d}a, & t < w(0), \\ \lambda_k \int_0^{w(\tau(t))} G_k(a)\mathrm{d}a, & t \ge w(0). \end{cases} \tag{30}$$

We have chosen a as the variable of integration to suggest the interpretation age, or time in system. This provides a simple interpretation of this formula. At time $t < w(0)$, none of the fluid that entered the system after time zero has been fully processed. So the integral term only needs to address departures resulting from expiring deadlines. Then, fluid arriving in $(0, t]$ remains in the system at time t if and only if the initial deadline exceeds the current age. Hence the simple form of the integral. At time $t \ge w(0)$, all fluid initially in the system has departed, and $w(\tau(t))$ can be interpreted as the age of the fluid departing the system via service completion at time t. So, for $t \ge \tau(t)$, $w(\tau(t))$ can be thought of as the age of the oldest fluid in the system at time t.

Nonabandoning Jobs Queue-Length Vector Fluid Approximation. Let $U = \{(w, p) \in \mathbb{R}_+^2 : w < p\}$. Then the mass present in U at time t is associated with fluid that doesn't abandon. For $1 \le k \le K$ and $t \in [0, \infty)$,

$$n_k(t) \equiv \zeta_k(t)(U) = \begin{cases} \zeta_k(0)(U_t) + \lambda_k \int_0^t G_k(w(v))\mathrm{d}v, & t < w(0), \\ \lambda_k \int_{\tau(t)}^t G_k(w(v))\mathrm{d}v, & t \ge w(0). \end{cases} \tag{31}$$

Here the variable of integration should be regarded as time, and the formula has the following interpretation. Fluid that has not departed the system by time t will not abandon prior to service completion if the initial deadline exceeds the fluid workload at the time of arrival.

For $t \ge w(0)$, we can rewrite this result in an alternative form that highlights the independence of the service time distributions to the deadline and interarrival time distributions. For $1 \le k \le K$ and $t \ge w(0)$, let

$$w_k(t) = \rho_k \int_{\tau(t)}^t G_k(w(v))\mathrm{d}v.$$

Then $w_k(t)$ denotes the amount of fluid workload in the system at time t due to class k fluid. For each $1 \le k \le K$, we have for all $t \ge w(0)$,

$$n_k(t) = \mu_k w_k(t).$$

In this form, the formula is reminiscent of linear state space collapse, but the dependence of $w_k(\cdot)$ on $w(\cdot)$ is nonlinear.

Abandoning Jobs Queue-Length Vector Fluid Approximation. Let $L = \{(w,p) \in \mathbb{R}_+^2 : p \leq w\}$. Then the mass present in L at time t is associated with fluid that abandons. For $1 \leq k \leq K$ and $t \in [0,\infty)$, let $a_k(t) = \zeta_k(t)(L)$. Then, for $1 \leq k \leq K$ and $t \in [0,\infty)$,

$$(32) \qquad a_k(t) = \begin{cases} \zeta_k(0)(L_t) + \lambda_k \int_0^t (G_k(t-v) - G_k(w(v)))\,\mathrm{d}v, & t < w(0), \\ \lambda_k \int_{\tau(t)}^t (G_k(t-v) - G_k(w(v)))\,\mathrm{d}v, & t \geq w(0). \end{cases}$$

Note that it is easy to verify that $z_k(t) = a_k(t) + n_k(t)$ for $1 \leq k \leq K$ and $t \in [0,\infty)$.

Age Related Fluid Approximations. Given $\vartheta \in \mathbf{I}$, let $\zeta(\cdot)$ denote the unique fluid model solution with $\zeta(0) = \vartheta$ and let $w(\cdot)$ denote the unique workload fluid model solution with $w(0) = w_\vartheta$. Fluid in the system at time $t > 0$ that arrived at time $s \in (0,t]$ is $t-s$ units old at time t. Hence, for fluid in the system at time $t > 0$ that arrives in $(0,t]$ to be of age at least u, it must have arrived by time $t-u$. Then $s \in (0, t-u]$. This fluid has residual offered waiting time $(w(s) - (t-s))^+$. Since it is in the system at time t, $w(s) - (t-s) > 0$, i.e., $w(s) + s > t$ so that $s \in (\tau(t), t-u]$. For $t \geq w(0)$, this is an empty time interval if $w(\tau(t)) \leq u$. Otherwise, for $t \geq w(0)$ and $s \in (\tau(t), t-u]$, $w(s) - (t-s) \in (0, w(t-u) - u]$. For $t \geq w(0)$ and $0 \leq u \leq t$, let

$$H(t,u) = [0, (w(t-u) - u)^+] \times \mathbb{R}_+.$$

This is the line $\{0\} \times \mathbb{R}_+$ if $u \geq w(\tau(t))$ and is a vertical stripe otherwise. Then, for $1 \leq k \leq K$, $t \geq w(0)$, and $0 \leq u \leq t$, we can interpret

$$z_k(t,u) \equiv \zeta_k(t)(H(t,u)),$$

to be the amount of class k fluid in the system at time t of age at least u.

If $t < w(0)$, some initial fluid may remain in the system at time t. We regard that fluid as being t units old. At time t, the residual offered waiting time of such fluid lies in $[0, w(0) - t]$. If we consider a time $0 \leq u \leq t < w(0)$ and ask for the total amount of fluid of age at least u, this would also include fluid that arrived after time zero and by time $t-u$. By (16), $w(t-u) - u > w(0) - t$. Hence the definition of $H(t,u)$ and interpretation of $z_k(t,u)$ naturally extend to all $0 \leq u \leq t < w(0)$. Then, for $1 \leq k \leq K$ we have the following: for $0 \leq u \leq t < w(0)$,

$$(33) \qquad z_k(t,u) = \zeta_k(0)(H(t,u)_t) + \lambda_k \int_u^t G_k(v)\mathrm{d}v,$$

and for $t \geq w(0)$,

$$z_k(t,u) = \begin{cases} \lambda_k \int_u^{w(\tau(t))} G_k(v)\mathrm{d}v, & 0 \leq u < w(\tau(t)), \\ 0, & w(\tau(t)) \leq u \leq t. \end{cases} \tag{34}$$

One can obtain similar approximations for the abandoning and nonabandoning queue length of a certain age or older by computing the measure of an appropriately chosen set.

Invariant States. An *invariant state* is a measure $\theta \in \mathbf{I}$ such that the unique fluid model solution $\zeta(\cdot)$ with initial measure θ satisfies $\zeta(t) = \theta$ for all $t \in [0, \infty)$. Here we identify the collection of invariant states.

For this, we begin by recalling some measure theoretic background. Details can be found in [6, Chapter 1]. Let

$$\mathcal{R} = \left\{ [a,b) \times [c,d) \subset \mathbb{R}_+^2 : a, c \in \mathbb{R}_+, a \leq b \leq \infty \text{ and } c \leq d \leq \infty \right\}.$$

Then $\mathcal{R}$ is an elementary family (a collection of sets that contains the emptyset, is closed under pairwise intersection, and such that complements of members of the collection can be written as finite unions of members of the collection). Further note that $\sigma(\mathcal{R}) = \mathcal{B}_2$. Let

$$\mathcal{R}' = \left\{ \cup_{l=1}^m R_l : m \in \mathbb{N} \text{ and } R_l \in \mathcal{R} \right\}.$$

Then $\mathcal{R}'$ is an algebra (a collection of sets that contains the emptyset and is closed under pairwise union and relative complementation). In fact, $\mathcal{R}'$ is the algebra generated by $\mathcal{R}$. It is easy to see that any finite pre-measure (additive $\mathbb{R}_+$ valued function that assigns value zero to the emptyset) on $\mathcal{R}$ extends to a finite pre-measure on $\mathcal{R}'$. Then, by the Carathéodory extension theorem [6, Theorem 1.14], any finite pre-measure defined on $\mathcal{R}'$ uniquely extends to a finite Borel measure on $\mathbb{R}_+^2$. Thus, in order to uniquely specify a finite Borel measure on $\mathbb{R}_+^2$, it suffices to specify a finite pre-measure on $\mathcal{R}$.

Given $w_l \leq w \leq w_u$, let $\theta^w \in \mathbf{I}$ be the unique finite Borel measure on $\mathbb{R}_+^2$ that for $1 \leq k \leq K$ satisfies

$$\theta_k^w([w,\infty) \times \mathbb{R}_+) = 0,$$

and for $0 \leq a < b \leq w$ and $0 \leq c < d \leq \infty \in \mathbb{R}_+$,

$$\theta_k^w([a,b) \times [c,d)) = \lambda_k \int_{w-b}^{w-a} \Gamma_k([c+u, d+u))\mathrm{d}u.$$

It is easy to verify that these relationships determine a finite pre-measure on $\mathcal{R}$. Indeed, since $[w,\infty) \times \mathbb{R}_+$ has measure zero, it suffices to specify each θ_k^w on sets in $\mathcal{R}$ that don't meet $[w,\infty) \times \mathbb{R}_+$. Define

$$\mathbf{J} = \{\theta^w : w_l \leq w \leq w_u\}.$$

THEOREM 3.3. *The set of invariant states is given by* $\mathbf{J}$.

The proof of Theorem 3.3 is given in Section 4. Note that for $w_l \le w \le w_u$ and $1 \le k \le K$, the fluid approximations for queue length z_k^w and nonabandoning queue length n_k^w take the form

$$\begin{aligned} z_k^w &\equiv \theta_k^w(\mathbb{R}_+^2) = \lambda_k \int_0^w G_k(u)\mathrm{d}u, \\ n_k^w &\equiv \theta_k^w(U) = \lambda_k w G_k(w). \end{aligned}$$

4. Properties of Fluid Model Solutions. In this section, we prove Theorems 3.1 and 3.3. The proof of existence and uniqueness for Theorem 3.1 is relatively straightforward. Indeed, since the right hand side of (26) only depends on (λ, μ, Γ) and ϑ, (26) can be regarded as a definition, provided that the integral term is a well defined function taking values in $\mathbf{M}_2^K$. In this regard, note that $\vartheta_k \in \mathbf{M}_2$ and $\delta^+_{w(s)} \times \Gamma_k \in \mathbf{M}_2$ for all $s \in [0,\infty)$ and $1 \le k \le K$. Hence, the main issue is to show that the integral is well defined on all of $\mathcal{B}_2$, which is demonstrated here using the Carathéodory extension theorem and Dynkin's $\pi\lambda$-theorem.

PROOF OF EXISTENCE AND UNIQUENESS FOR THEOREM 3.1. Fix $\vartheta \in \mathbf{I}$. First we verify existence of a fluid model solution with initial measure ϑ. For this, fix $1 \le k \le K$ and $0 < t < \infty$. Given $B \in \mathcal{R}$, we have that $B = [a,b) \times [c,d)$ for some $a, c \in \mathbb{R}_+$, $a \le b \le \infty$, and $c \le d \le \infty$. Define $f : [0,t] \to [0,1]$ by

$$f(s) = \left(\delta^+_{w(s)} \times \Gamma_k\right)(B_{t-s}).$$

By (17), the fact that $w_l > 0$ and monotonicity properties of $w(\cdot)$, $w(s) > 0$ for all $s > 0$. Then, since $t > 0$, we have that for $s \in (0,t]$,

$$f(s) = \left(\delta_{w(s)} \times \Gamma_k\right)(B_{t-s}).$$

Therefore, for $s \in [0,t]$,

$$f(s) = \begin{cases} G_k(c+t-s) - G_k(d+t-s), & \text{if } a+t-s \le w(s) < b+t-s, \\ 0, & \text{otherwise.} \end{cases}$$

We see that $a+t-s \le w(s) < b+t-s$ if and only if $a+t \le w(s)+s < b+t$ if and only if $\tau(a+t) \le s < \tau(b+t)$, where $\tau(\infty) = \infty$. Hence, for $s \in [0,t]$,

$$f(s) = \begin{cases} G_k(c+t-s) - G_k(d+t-s), & \text{if } \tau(a+t) \le s < \tau(b+t), \\ 0, & \text{otherwise.} \end{cases}$$

Then, since $G_k(\cdot)$ and $\tau(\cdot)$ are continuous, f is Borel measurable and $\int_0^t f(s)ds$ is well defined. Further, since $\tau(\cdot)$ is monotone increasing,

$$\int_0^t f(s)\mathrm{d}s = \int_{t\wedge\tau(a+t)}^{t\wedge\tau(b+t)} \left(G_k(c+t-s) - G_k(d+t-s)\right)\mathrm{d}s.$$

For $B \in \mathcal{R}$, define

$$\gamma_k(t)(B) = \vartheta_k(B_t) + \lambda_k \int_0^t \left(\delta^+_{w(s)} \times \Gamma_k\right)(B_{t-s})\mathrm{d}s.$$

It is clear that $\gamma_k(t)(\emptyset) = 0$. Further, $\gamma_k(t)(\mathbb{R}^2_+) \le \vartheta_k(\mathbb{R}^2_+) + \lambda_k t < \infty$. So, in order to verify that $\gamma_k(t)$ is a finite premeasure on $\mathcal{R}$, it suffices to verify countable additivity. This amounts to demonstrating that the summation and integral can be interchanged, which is an immediate consequence of the monotone convergence theorem. Then, by the Carathéodory extension theorem, $\gamma_k(t)$ extends to a finite Borel measure $\gamma^*_k(t)$ on $\mathcal{B}_2$.

We must verify that $\gamma^*_k(t)$ satisfies (26) for all $B \in \mathcal{B}_2$. For this, we use Dynkin's $\pi\lambda$-theorem [4, Chapter 1 Theorem 3.3]. Let

$$\mathcal{P} = \{[a,\infty) \times [c,\infty) : 0 \le a, c < \infty\}. \tag{35}$$

This is a π-system since it is closed under intersection. Let

$$\mathcal{L} = \left\{ B \in \mathcal{B}_2 : \gamma^*_k(t)(B) = \vartheta_k(B_t) + \lambda_k \int_0^t \left(\delta_{w(s)} \times \Gamma_k\right)(B_{t-s})\mathrm{d}s \right\}.$$

Then $\mathcal{L}$ is a λ-system since $\mathbb{R}^2_+ \in \mathcal{L}$, $\mathcal{L}$ is closed under countable unions (by the monotone convergence theorem), and $A \setminus B \in \mathcal{L}$ whenever $B, A \in \mathcal{L}$ and $B \subset A$. Further,

$$\mathcal{P} \subset \mathcal{R} \subset \mathcal{L} \subset \mathcal{B}_2.$$

Then, since the σ-algebra generated by $\mathcal{P}$ is $\mathcal{B}_2$, it follows from Dynkin's $\pi\lambda$-theorem that $\mathcal{L} = \mathcal{B}_2$.

Since $1 \le k \le K$ and $t > 0$ were arbitrary, it follows that $\gamma^* : [0,\infty) \to \mathbf{M}_2^K$, which is given by

$$\gamma^*(t) = (\gamma^*_1(t), \dots, \gamma^*_K(t)),$$

is a fluid model solution. Uniqueness is immediate since any fluid model solution ζ such that $\zeta(0) = \vartheta$ satisfies $\zeta_k(t)(B) = \gamma^*_k(t)(B)$ for all $B \in \mathcal{R}$ (see [6, Theorem 1.14]). $\square$

PROOF OF PROPERTY 1 FOR THEOREM 3.1. Fix $\vartheta \in \mathbf{I}$. Let $w(\cdot)$ be the unique workload fluid model solution with $w(0) = w_\vartheta$ and $\zeta(\cdot)$ be the unique fluid model solution with $\zeta(0) = \vartheta$. Since $\zeta(0) = \vartheta$ and $w(0) = w_\vartheta$, $w_{\zeta(0)} = w(0)$. Fix $t \in (0,\infty)$ and $\varepsilon \in (0, w(t))$. Set $B = [w(t) - \varepsilon, \infty) \times \mathbb{R}_+$. Then $B_{(2\varepsilon,0)} = [w(t) + \varepsilon, \infty) \times \mathbb{R}_+$. We wish to show that $\zeta_+(t)(B) > 0$ and $\zeta_+(t)(B_{(2\varepsilon,0)}) = 0$. Then, since $\varepsilon \in (0, w(t))$ is arbitrary, it follows that $w_{\zeta(t)} = w(t)$.

Suppose that $\zeta_+(t)(B_{(2\varepsilon,0)}) > 0$. Then by (26), there exists $s \in [0,t]$ such that $w(t) + \varepsilon + t - s \le w(s)$, i.e., $w(t) + t + \varepsilon \le w(s) + s$. But $w(\cdot) + \iota(\cdot)$ is strictly increasing, so that $w(t) + t + \varepsilon > w(s) + s$, which is a contradiction. Thus, $\zeta_+(t)(B_{(2\varepsilon,0)}) = 0$.

Since $w(\cdot)+\iota(\cdot)$ is continuous and strictly increasing, there exists $s_1 \in [0,t)$ such that $w(t) + t - \varepsilon \le w(s_1) + s_1$. Then, for all $s \in (s_1, t]$, $w(t) + t - \varepsilon \le w(s) + s$. Let $s_2 = (t - d_{\max})^+$. Then $s_2 \in [0,t)$ and $t - s < d_{\max}$ for all $s \in (s_2, t]$. Let $s^* = s_1 \vee s_2$. Then $s^* < t$ and

$$\sum_{k=1}^{K} \int_0^t \left(\delta^+_{w(s)} \times \Gamma_k\right)(B_{t-s})\mathrm{d}s \ge \sum_{k=1}^{K} \int_{s_1}^t G_k(t-s)\mathrm{d}s \ge \sum_{k=1}^{K} \int_{s^*}^t G_k(t-s)\mathrm{d}s > 0.$$

Hence, by (26), $\zeta_+(t)(B) > 0$. □

The next goal is to verify Properties 2 and 3 for Theorem 3.1. For this, we state and prove the following lemma. Property 2 follows immediately from Lemma 4.1 by letting ε decrease to zero. To verify Property 3, we must show that fluid model solutions change very little over short time intervals. Note that there are three mechanisms that cause the measure-valued function to change: fluid arriving, fluid departing, and the measure evolving. Verifying that the fluid departs in a smooth way is the main issue that needs to be addressed. We will use the result in the following lemma to assist with this as well.

LEMMA 4.1. *Given $\vartheta \in \mathbf{I}$, let $\zeta(\cdot)$ be the unique fluid model solution such that $\zeta(0) = \vartheta$. For every $T, \varepsilon > 0$, there exists $\kappa > 0$ such that*

$$\max_{1\le k\le K} \sup_{t\in[0,T]} \sup_{x,y\in\mathbb{R}_+} \zeta_k(t)(C^\kappa_{(x,y)}) < \varepsilon.$$

To ease the reader's efforts to follow the proof of Lemma 4.1 given below, we outline the basic strategy. Having a good understanding of this deterministic argument will assist the reader in following the stochastic generalization used to prove Lemma 5.7. The basic idea in this case is to use (26) on a given

κ enlargement of a corner set. Then (24) can be used to bound the contribution from mass present in the system at time zero. Next one determines the time interval during which mass must arrive in order to contribute to the vertical portion of the enlarged corner set. As the reader will see, the end points of this time interval can be expressed in terms of the function $\tau(\cdot)$. Then one obtains a bound on the mass that can be present in the horizontal portion of the enlarged corner set. This second part turns out to be fairly easy to bound, as the reader will see. Hence the main difficultly is to bound the amount of mass that falls in the vertical portion. This relies on bounding the length of the aforementioned time interval, which can be done by demonstrating that the function $w(\cdot)+\iota(\cdot)$ increases sufficiently quickly. But $w(\cdot)+\iota(\cdot)$ actually increases quite slowly at times when the workload fluid model solution is near $d_{\max}$, i.e., possibly at small times. So one must wait a short amount of time (until time δ in the proof) before implementing this strategy during which just a small amount of mass enters the entire system. Once that small amount of time has elapsed, (16) can be used to obtain the desired bound. The mathematical details are given next.

PROOF OF LEMMA 4.1. Fix $T,\varepsilon>0$. Set $\lambda_+=\sum_{k=1}^K\lambda_k$, $\delta=\varepsilon/(4\lambda_+)$, $M=w_u\vee w(\delta)$, and $c=\sum_{k=1}^K\rho_kG_k(M)$. Note that $c>0$ since $M<d_{\max}$. Further, by monotonicity properties of workload fluid model solutions, $w(u)\le M$ for all $u\ge\delta$.

For $t\in[0,T]$, $1\le k\le K$, $x,y\in\mathbb{R}_+$, and $\kappa>0$, by (26),

$$\zeta_k(t)\left(C^\kappa_{(x,y)}\right)=\zeta_k(0)\left(\left(C^\kappa_{(x,y)}\right)_t\right)+\lambda_k\int_0^t\left(\delta^+_{w(s)}\times\Gamma_k\right)\left(\left(C^\kappa_{(x,y)}\right)_{t-s}\right)\mathrm{d}s.$$

By (24) there exists κ_0 such that for all $0<\kappa<\kappa_0$,

$$\max_{1\le k\le K}\sup_{t\in[0,T]}\sup_{x,y\in\mathbb{R}_+}\zeta_k(0)\left(\left(C^\kappa_{(x,y)}\right)_t\right)<\frac{\varepsilon}{4}.$$

Hence, for $t\in[0,T]$, $1\le k\le K$, $x,y\in\mathbb{R}_+$, and $0<\kappa<\kappa_0$,

$$\zeta_k(t)\left(C^\kappa_{(x,y)}\right)<\frac{\varepsilon}{4}+\lambda_k\int_0^t\left(\delta^+_{w(s)}\times\Gamma_k\right)\left(\left(C^\kappa_{(x,y)}\right)_{t-s}\right)\mathrm{d}s. \tag{36}$$

We must show that there exists $\kappa^*\le\kappa_0$ such that for all $t\in[0,T]$, $1\le k\le K$, $x,y\in\mathbb{R}_+$, and $0<\kappa<\kappa^*$, the integral term is bounded above by $3\varepsilon/4$.

Fix $t\in[0,T]$, $1\le k\le K$, and $x,y\in\mathbb{R}_+$. For $s\in[0,T]$ and $\kappa>0$, let

$$h_k(s,\kappa)=G_k((y-\kappa)^++t-s)-G_k(y+\kappa+t-s).$$

For $0 < \kappa < \kappa_0$, the integrand in (36) at time $s \in [0, t]$ is bounded above by

$$\begin{cases} 0, & \text{if } w(s) + s < (x-\kappa)^+ + t, \\ 1, & \text{if } (x-\kappa)^+ + t \le w(s) + s \le x + \kappa + t, \\ h_k(s,\kappa), & \text{if } x + \kappa + t < w(s) + s. \end{cases}$$

Equivalently, for $0 < \kappa < \kappa_0$, using continuity and monotonicity properties of $w(\cdot) + \iota(\cdot)$, the integrand in (36) at time $s \in [0, t]$ is bounded above by

$$\begin{cases} 0, & \text{if } 0 \le s < \tau((x-\kappa)^+ + t), \\ 1, & \text{if } \tau((x-\kappa)^+ + t) \le s \le \tau(x + \kappa + t), \\ h_k(s,\kappa), & \text{if } \tau(x + \kappa + t) < s \le t. \end{cases} \tag{37}$$

This together with (36) yields that for $0 < \kappa < \kappa_0$,

$$\begin{aligned} \zeta_k(t)\left(C^{\kappa}_{(x,y)}\right) &< \frac{\varepsilon}{4} + \lambda_k \int_{\tau((x-\kappa)^+ + t)\wedge t}^{\tau(x+\kappa+t)\wedge t} \mathrm{d}s + \lambda_k \int_{\tau(x+\kappa+t)\wedge t}^{t} h_k(s,\kappa)\mathrm{d}s \\ &\le \frac{\varepsilon}{4} + \lambda_k \left(\tau(x+\kappa+t)\wedge t - \tau((x-\kappa)^+ + t)\wedge t\right) \\ &\quad + \lambda_k \int_{(y-\kappa)^+}^{y+\kappa} G_k(u)\mathrm{d}u \\ &\le \frac{\varepsilon}{4} + \lambda_k \left(\tau(x+\kappa+t)\wedge t - \tau((x-\kappa)^+ + t)\wedge t\right) + 2\lambda_k\kappa. \end{aligned} \tag{38}$$

For $0 < \kappa < \kappa_0$, let

$$\Delta(\kappa) = \tau(x+\kappa+t)\wedge t - \tau((x-\kappa)^+ + t)\wedge t.$$

Fix $0 < \kappa < \kappa_0$. If $\tau((x-\kappa)^+ + t) \ge t$, $\Delta(\kappa) = 0$. Also, if $x+\kappa+t \le w(0)$, then $\Delta(\kappa) = 0$. Henceforth, we assume that $\tau((x-\kappa)^+ + t) < t$ and $x+\kappa+t > w(0)$. If $\tau(x+\kappa+t)\wedge t \le \delta$, $\Delta(\kappa) \le \delta$. Otherwise, $\tau(x+\kappa+t)\wedge t > \delta$, and there are two cases to consider. First consider the case where $\tau((x-\kappa)^+ + t) \ge \delta$. Then $(x-\kappa)^+ + t > w(0)$ since $\tau((x-\kappa)^+ + t) > 0$. Hence, by (29) and (16), since $\tau((x-\kappa)^+ + t) \ge \delta$,

$$\begin{aligned} 2\kappa &= x + \kappa + t - (x - \kappa + t) \\ &\ge x + \kappa + t - ((x-\kappa)^+ + t) \\ &= w(\tau(x+\kappa+t)) + \tau(x+\kappa+t) \\ &\quad - w(\tau((x-\kappa)^+ + t)) - \tau((x-\kappa)^+ + t) \\ &= \sum_{k=1}^{K} \rho_k \int_{\tau((x-\kappa)^+ + t)}^{\tau(x+\kappa+t)} G_k(w(u))\mathrm{d}u \end{aligned}$$

$$\begin{aligned} &\geq \sum_{k=1}^{K} \rho_k \int_{\tau((x-\kappa)^+ +t)}^{\tau(x+\kappa+t)} G_k(M)\mathrm{d}u \\ &\geq c\Delta(\kappa). \end{aligned}$$

In particular, $\Delta(\kappa) \leq 2\kappa/c$. The other case to consider is $\tau((x-\kappa)^+ + t) < \delta$. Then

$$\Delta(\kappa) \leq \tau(x+\kappa+t) \wedge t = (\tau(x+\kappa+t) \wedge t - \delta) + \delta.$$

Further, by (29) and monotoncity of $w(\cdot) + \iota(\cdot)$,

$$\begin{aligned} x - \kappa + t &\leq (x-\kappa)^+ + t \leq w(\tau((x-\kappa)^+ + t)) + \tau((x-\kappa)^+ + t) \\ &\leq w(\delta) + \delta \leq w(\tau(x+\kappa+t)) + \tau(x+\kappa+t) = x+\kappa+t. \end{aligned}$$

Then, by (29) and (16),

$$\begin{aligned} 2\kappa &\geq x+\kappa+t - w(\delta) - \delta \\ &= w(\tau(x+\kappa+t)) + \tau(x+\kappa+t) - w(\delta) - \delta \\ &= \sum_{k=1}^{K} \rho_k \int_{\delta}^{\tau(x+\kappa+t)} G_k(w(u))\mathrm{d}u \\ &\geq \sum_{k=1}^{K} \rho_k \int_{\delta}^{\tau(x+\kappa+t)} G_k(M)\mathrm{d}u \\ &= c\left(\tau(x+\kappa+t) - \delta\right). \end{aligned}$$

In particular,

$$\Delta(\kappa) \leq \frac{2\kappa}{c} + \delta, \tag{39}$$

which is the largest of the four upper bounds.

By combining (38), (39), and the definition of δ it follows that for $0 < \kappa < \kappa_0$,

$$\zeta_k(t)\left(C^{\kappa}_{(x,y)}\right) < \frac{\varepsilon}{4} + \frac{2\lambda_k\kappa}{c} + \lambda_k\delta + 2\lambda_k\kappa \leq \frac{\varepsilon}{2} + \frac{2\lambda_k\kappa}{c} + 2\lambda_k\kappa.$$

Let $0 < \kappa^* \leq \kappa_0$ be such that for all $0 < \kappa < \kappa^*$

$$\max_{1\leq k\leq K} 2\lambda_k\left(\frac{1}{c} + 1\right)\kappa < \frac{\varepsilon}{2}.$$

Then, for all $0 < \kappa < \kappa^*$, $\zeta_k(t)(C^{\kappa}_{(x,y)}) < \varepsilon$. Since $1 \leq k \leq K$, $t \in [0,T]$, and $x, y, \in \mathbb{R}_+$ were chosen arbitrarily, the result follows. $\square$

Next, Lemma 4.1 is used to prove continuity for Theorem 3.1. For this, we need to specify a metric on $\mathbf{M}_2$ that induces the topology of weak convergence. For efficiency, we will specify such a metric on $\mathbf{M}_i$ for $i = 1, 2$. Given $i = 1, 2$ and ζ, $\zeta' \in \mathbf{M}_i$, let $\mathbf{d}[\zeta, \zeta']$ denote the Prohorov distance between ζ and ζ'. Specifically,

$$\begin{aligned}\mathbf{d}[\zeta,\zeta'] \;&=\; \inf\{\varepsilon > 0 : \zeta(B) \le \zeta'(B^\varepsilon) + \varepsilon \text{ and } \zeta'(B) \le \zeta(B^\varepsilon) + \varepsilon \\ &\qquad \text{for all closed } B \in \mathcal{B}_i\}.\end{aligned}$$

Then the following is a natural metric on $\mathbf{M}_i^K$. Given $i = 1, 2$ and $\zeta, \zeta' \in \mathbf{M}_i^K$, define

$$\mathbf{d}_K[\zeta,\zeta'] = \max_{1\le k\le K} \mathbf{d}[\zeta_k,\zeta'_k]. \tag{40}$$

In the following proof, we show that for each fluid model solution $\zeta(\cdot)$, $\lim_{h\to 0} \mathbf{d}_K[\zeta(t+h), \zeta(t)] = 0$ for all $t \in [0,\infty)$.

Proof of Property 3 for Theorem 3.1. Let $\vartheta \in \mathbf{I}$ and let $\zeta(\cdot)$ be the unique fluid model solution such that $\zeta(0) = \vartheta$. It suffices to prove continuity of $\zeta_k(\cdot)$ on $[0,\infty)$ for each $1 \le k \le K$. Fix $1 \le k \le K$. We prove continuity of $\zeta_k(\cdot)$ on $[0, T]$ for each $T > 0$. Fix $T, \varepsilon > 0$. By Lemma 4.1, there exists κ such that for all $0 < h < \kappa$,

$$\sup_{s\in[0,T]} \zeta_k(s)(C^h) < \varepsilon.$$

Let $0 \le s < t \le T$ be such that $t - s < \min(\kappa, \varepsilon/\max(2, \lambda_k))$. Note that $(\delta^+_{w(u)} \times \Gamma_k)(\mathbb{R}^2_+) = 1$ for all $u \in [0,\infty)$. Let $B \in \mathcal{B}_2$ be closed and set $h = t - s > 0$. Note that $B_h \subset B^{2h}$. By subtracting (26) applied to B_h at time s from (26) applied to B at time t, we obtain

$$\begin{aligned}\zeta_k(t)(B) \;&=\; \zeta_k(s)(B_h) + \lambda_k \int_s^t \left(\delta^+_{w(u)} \times \Gamma_k\right)(B_{t-u})\mathrm{d}u \\ &\le\; \zeta_k(s)(B^{2h}) + \lambda_k h \\ &<\; \zeta_k(s)(B^\varepsilon) + \varepsilon.\end{aligned}$$

Consider (26) applied to B at time s, to B^{2h} at time t, and to C^h at time s. For $u \in [0, s]$, we see that $(w, p) \in B_u$ implies that $(w-u, p-u) \in B$, which implies that either $(w-u, p-u) \in B \cap C^h \subset C^h$ or $(w-u, p-u) \in B \setminus C^h$. If $(w-u, p-u) \in C^h$, then $(w, p) \in (C^h)_u$. If $(w-u, p-u) \in B \setminus C^h$, then $(w-u-h, p-u-h) \in B^{2h}$ and so $(w, p) \in (B^{2h})_{u+h}$ and $u + h =$

$u+t-s \le t$. Therefore, for $u \in [0,s]$, $B_{s-u} \subset (C^h)_{s-u} \cup (B^{2h})_{t-v}$, where $v = t-(s-u+h) \in [0,t]$. Hence,

$$\zeta_k(s)(B) \le \zeta_k(t)(B^{2h}) + \zeta_k(s)(C^h) < \zeta_k(t)(B^\varepsilon) + \varepsilon.$$

Then for all $0 \le s < t \le T$ such that $0 < t-s < \min(\kappa, \varepsilon/\max(2,\lambda_k))$,

$$\mathbf{d}[\zeta_k(t), \zeta_k(s)] < \varepsilon.$$

Hence, $\zeta_k(\cdot)$ is continuous on $[0,T]$. Since $T > 0$ was arbitrary, $\zeta_k(\cdot)$ is continuous on $[0,\infty)$. Since $1 \le k \le K$ was arbitrary, $\zeta(\cdot)$ is continuous on $[0,\infty)$. □

Now that each statement in Theorem 3.1 has been verified, we prove Theorem 3.3.

Proof of Theorem 3.3. First suppose that $\vartheta \in \mathbf{I}$ is an invariant state. We must show that $\vartheta \in \mathbf{J}$, i.e., we must show that for some $w \in [w_l, w_u]$, $\vartheta_k(B) = \theta_k^w(B)$ for all $B \in \mathcal{B}_2$ and $1 \le k \le K$. Since ϑ is an invariant state, Theorem 3.1 Property 1 implies that the unique workload fluid model solution such that $w(0) = w_\vartheta$ is constant, i.e., $w(t) = w_\vartheta$ for all $t \in [0,\infty)$. Then, by the monotonicity properties of workload fluid model solutions, $w_l \le w_\vartheta \le w_u$. Let $w = w_\vartheta$. Fix $1 \le k \le K$, $B \in \mathcal{P}$, and $t > w$. Then $B = [a,\infty) \times [c,\infty)$ for some $a, c \in \mathbb{R}_+$. Further, $a+t-w > 0$ and $(w-a)^+ < t$. Hence, by (26), we have

$$\begin{aligned}
\vartheta_k(B) &= \lambda_k \int_0^t \left(\delta_w^+ \times \Gamma_k\right)(B_{t-s})\mathrm{d}s \\
&= \lambda_k \int_{(a+t-w)\wedge t}^t G_k(c+t-s)\mathrm{d}s \\
&= \lambda_k \int_0^{(w-a)^+} G_k(c+u)\mathrm{d}u \\
&= \theta_k^w(B).
\end{aligned}$$

Since $B \in \mathcal{P}$ was arbitrary, ϑ_k and θ_k^w agree on $\mathcal{P}$. Since ϑ_k and θ_k^w agree on $\mathcal{P}$, they agree on $\mathcal{R}$ and therefore they agree on $\mathcal{B}_2$. So $\vartheta_k = \theta_k^w$. Since $1 \le k \le K$ was arbirary, $\vartheta = \theta^w \in \mathbf{J}$.

Next, fix $w_l \le w \le w_u$. Then the unique fluid workload solution $w(\cdot)$ with $w(0) = w$ is constant, i.e., $w(t) = w$ for all $t \in [0,\infty)$. Set $\zeta(t) = \theta^w$ for all $t \in [0,\infty)$. Then, in order to show that $\zeta(\cdot)$ is a fluid model solution, it suffices to verify that $\zeta(\cdot)$ satisfies (26). For this, it suffices to verify that for all $B \in \mathcal{P}$, $1 \le k \le K$, and $t > 0$,

$$\theta_k^w(B) = \theta_k^w(B_t) + \lambda_k \int_0^t \left(\delta_w^+ \times \Gamma_k\right)(B_{t-s})\mathrm{d}s.$$

For this fix $B \in \mathcal{P}$, $1 \le k \le K$, and $t > 0$. We have

$$\begin{aligned} \theta_k^w(B_t) \quad &+ \quad \lambda_k \int_0^t \left(\delta_w^+ \times \Gamma_k\right)(B_{t-s})\mathrm{d}s \\ &= \quad \theta_k^w(B_t) + \lambda_k \int_{(a+t-w)^+ \wedge t}^t G_k(c+t-s)\mathrm{d}s. \end{aligned}$$

Case 1: Suppose that $a \ge w$. Then $(a+t-w)^+ \wedge t = t$ and it follows that

$$\theta_k^w(B_t) + \lambda_k \int_0^t \left(\delta_w^+ \times \Gamma_k\right)(B_{t-s})\mathrm{d}s = 0 = \theta_k^w(B).$$

Case 2: Suppose that $a < w \le a+t$. Then $(a+t-w)^+ \wedge t = a+t-w$ and

$$\begin{aligned} \theta_k^w(B_t) \quad &+ \quad \lambda_k \int_0^t \left(\delta_w^+ \times \Gamma_k\right)(B_{t-s})\mathrm{d}s \\ &= \quad 0 + \lambda_k \int_{a+t-w}^t G_k(c+t-s)\mathrm{d}s \\ &= \quad \lambda_k \int_0^{w-a} G_k(c+u)\mathrm{d}u \\ &= \quad \theta_k^w(B). \end{aligned}$$

Case 3: If $a+t < w$, then $(a+t-w)^+ \wedge t = 0$ and

$$\begin{aligned} \theta_k^w(B_t) \quad &+ \quad \lambda_k \int_0^t \left(\delta_w^+ \times \Gamma_k\right)(B_{t-s})\mathrm{d}s \\ &= \quad \theta_k^w([a+t,w) \times [c+t,\infty)) + \lambda_k \int_0^t G_k(c+t-s)\mathrm{d}s \\ &= \quad \lambda_k \int_0^{w-a-t} G_k(c+t+u)\mathrm{d}u + \lambda_k \int_0^t G_k(c+u)\mathrm{d}u \\ &= \quad \lambda_k \int_t^{w-a} G_k(c+u)\mathrm{d}u + \lambda_k \int_0^t G_k(c+u)\mathrm{d}u \\ &= \quad \lambda_k \int_0^{w-a} G_k(c+u)\mathrm{d}u \\ &= \quad \theta_k^w(B). \end{aligned}$$

Hence θ^w is an invariant state. $\square$

5. Proof of Tightness. In this section, we prove that the sequence $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$ is relatively compact under the standing assumption (27). For this, we apply [5, Corollary 3.7.4]. In particular, it suffices to prove compact containment and an oscillation inequality. This is done in Lemmas 5.5 and 5.11 below. Throughout,

$$\lambda_+ = \sum_{k=1}^K \lambda_k \qquad \text{and} \qquad g_+ = \sum_{k=1}^K \frac{1}{\gamma_k}.$$

5.1. *Preliminaries.* For $1 \le k \le K$ and $t \in [0,\infty)$, let $\lambda_k^*(t) = \lambda_k t$. For $1 \le k \le K$, let $\mathcal{D}_k^*(\cdot) = \lambda_k^*(\cdot)\Gamma_k$. Define $\lambda^*(\cdot) = (\lambda_1^*(\cdot),\dots,\lambda_K^*(\cdot))$ and $\mathcal{D}^*(\cdot) = (\mathcal{D}_1^*(\cdot)\dots,\mathcal{D}_K^*(\cdot))$. By a functional law of large numbers for renewal processes, as $n\to\infty$,

$$\bar{E}^n(\cdot) \Rightarrow \lambda^*(\cdot). \tag{41}$$

Further, as a special case of Lemma 5.1 stated below, the following functional law of large numbers for the deadline process holds: as $n\to\infty$,

$$\left(\bar{\mathcal{D}}^n(\cdot), \left\langle \chi, \bar{\mathcal{D}}^n(\cdot)\right\rangle\right) \Rightarrow \left(\mathcal{D}^*(\cdot), \left\langle \chi, \mathcal{D}^*(\cdot)\right\rangle\right). \tag{42}$$

To state Lemma 5.1, we need to introduce some additional notation and asymptotic assumptions.

Asymptotic Assumptions (AA). For each $n\in\mathbb{N}$ suppose that we have K independent sequences of strictly positive independent and identically distributed random variables that are independent of the exogenous arrival process. For each $n\in\mathbb{N}$, denote the kth sequence by $\{g_{k,i}^n\}_{i\in\mathbb{N}}$, and for each $n\in\mathbb{N}$ and $1\le k\le K$, denote the distribution of $g_{k,1}^n$ by Γ_k^n. We assume that Γ_k^n has a finite mean $1/\gamma_k^n$, but we do not necessarily assume that Γ_k^n is continuous. We further assume that for each $1\le k\le K$,

$$\lim_{M\to\infty}\sup_{n\in\mathbb{N}} \left\langle \chi 1_{(M,\infty)}, \Gamma_k^n\right\rangle = 0, \tag{43}$$

and, as $n\to\infty$,

$$\Gamma_k^n \xrightarrow{w} \Gamma_k. \tag{44}$$

One interpretation of (AA) is that for large n and $1\le k\le K$, Γ_k^n approximates Γ_k. So, for $n\in\mathbb{N}$ and $1\le k\le K$, one can regard $\{g_{k,i}^n\}_{i\in\mathbb{N}}$ as a collection of approximate patience times. Here we are primarily interested in two particular choices of approximate patience times. These are introduced below, shortly after the statement of Lemma 5.3.

Note that (43) is a uniform integrability condition and, together with (44), it implies that for each $1 \le k \le K$, as $n \to \infty$,

$$1/\gamma_k^n \to 1/\gamma_k.$$

For each $n \in \mathbb{N}$, let $\Gamma^n = \Gamma_1^n \times \cdots \times \Gamma_K^n$ and let $\Gamma = \Gamma_1 \times \cdots \times \Gamma_K$. For $n \in \mathbb{N}$, $1 \le k \le K$, and $t \in [0, \infty)$, define

$$\mathcal{G}_k^n(t) = \sum_{i=1}^{E_k^n(t)} \delta^+_{g_{k,i}^n} \qquad \text{and} \quad \bar{\mathcal{G}}_k^n(t) = \frac{1}{n}\mathcal{G}_k^n(t).$$

For $n \in \mathbb{N}$, let $\mathcal{G}^n(\cdot) = (\mathcal{G}_1^n(\cdot), \dots, \mathcal{G}_K^n(\cdot))$ and $\bar{\mathcal{G}}^n(\cdot) = (\bar{\mathcal{G}}_1^n(\cdot), \dots, \bar{\mathcal{G}}_K^n(\cdot))$. So then, for each $n \in \mathbb{N}$, $\mathcal{G}^n(\cdot) \in \mathbf{D}([0, \infty), \mathbf{M}_1^K)$. We refer to $\mathcal{G}^n(\cdot)$, $n \in \mathbb{N}$, as a deadline related process.

LEMMA 5.1. *Suppose that (AA) holds. Then, as $n \to \infty$,*

$$\left(\bar{\mathcal{G}}^n(\cdot), \left\langle \chi, \bar{\mathcal{G}}^n(\cdot)\right\rangle\right) \Rightarrow \left(\mathcal{D}^*(\cdot), \langle \chi, \mathcal{D}^*(\cdot)\rangle\right).$$

Lemma 5.1 holds by (41) and [10, Theorem 5.1]. To see this, note the following. Respectively, the number of classes and deadline distributions play the same role for the deadline related processes defined here as the number of routes and service time distributions do for the load process defined in [10]. Then [10, Theorem 5.1] holds under [10, Assumption (A)], and the conditions in [10, Assumption (A)] relevant to [10, Theorem 5.1] are [10, (4.8)–(4.13)]. Conditions [10, (4.8)–(4.9)] hold here because the deadlines are assumed to be strictly positive and have finite means. Condition [10, (4.10)] corresponds to (41) above. Conditions [10, (4.11)–(4.13))] hold here due to (43) and (44) above.

For each $n \in \mathbb{N}$ and $1 \le k \le K$, define the fluid scaled increments as follows: for $0 \le s \le t < \infty$,

$$\bar{E}_k^n(s,t) = \bar{E}_k^n(t) - \bar{E}_k^n(s) \qquad \text{and} \qquad \lambda_k^*(s,t) = \lambda_k^*(t) - \lambda_k^*(s),$$

and

$$\bar{\mathcal{G}}_k^n(s,t) = \bar{\mathcal{G}}_k^n(t) - \bar{\mathcal{G}}_k^n(s) \qquad \text{and} \qquad \mathcal{D}_k^*(s,t) = \mathcal{D}_k^*(t) - \mathcal{D}_k^*(s).$$

Then, by (41), given $T, \varepsilon, \eta > 0$,

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{0\le s\le t\le T} \left|\bar{E}_k^n(s,t) - \lambda_k^*(s,t)\right| \le \varepsilon\right) \ge 1 - \eta. \tag{45}$$

Similarly, the following corollary holds as a consequence of Lemma 5.1 and continuity of the measures Γ_k, $1 \le k \le K$.

COROLLARY 5.2. *Let $T, \varepsilon, \eta > 0$ and $0 \le a \le b < \infty$. Suppose that (AA) holds. Then*

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{0\le s\le t\le T} \left| \bar{\mathcal{G}}_k^n(s,t)((a,b)) - \mathcal{D}_k^*(s,t)((a,b)) \right| \le \varepsilon \right) \ge 1-\eta.$$

We also wish to consider a version of a deadline related process involving residual approximate patience times. We refer to such a process as a residual deadline related process. To define these, let $\{g_{k,i}^n\}_{i\in\mathbb{N}}$, $1 \le k \le K$, and $n \in \mathbb{N}$, satisfy (AA). For each $n \in \mathbb{N}$, $1 \le k \le K$ and $t \in [0,\infty)$, let

$$g_{k,i}^n(t) = \left(g_{k,i}^n - \left(t - t_{k,Z_k^n(0)+i}^n \right)^+ \right)^+,$$

and set

$$\mathcal{R}_k^n(t) = \sum_{i=1}^{E_k^n(t)} \delta^+_{g_{k,i}^n(t)} \qquad \text{and} \qquad \bar{\mathcal{R}}_k^n(t) = \frac{\mathcal{R}_k^n(t)}{n}.$$

Then, for each $n \in \mathbb{N}$ and $1 \le k \le K$, $\mathcal{R}_k^n(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M}_1)$. For $n \in \mathbb{N}$, set $\mathcal{R}^n(\cdot) = (\mathcal{R}_1^n(\cdot), \ldots, \mathcal{R}_K^n(\cdot))$ and $\bar{\mathcal{R}}^n(\cdot) = (\bar{\mathcal{R}}_1^n(\cdot), \ldots, \bar{\mathcal{R}}_K^n(\cdot))$. For each $1 \le k \le K$ and $t \in [0,\infty)$, let $\mathcal{R}_k^*(t) \in \mathbf{M}_1$ be the measure that is absolutely continuous with respect to Lebesgue measure with density $\lambda_k(G_k(\cdot) - G_k(\cdot + t))$. Set $\mathcal{R}^*(\cdot) = (\mathcal{R}_1^*(\cdot), \ldots, \mathcal{R}_K^*(\cdot))$.

LEMMA 5.3. *Suppose that (AA) holds. Then, as $n \to \infty$,*

$$\bar{\mathcal{R}}^n(\cdot) \Rightarrow \mathcal{R}^*(\cdot).$$

In order to prove Lemma 5.3, we must first prove that $\{\bar{\mathcal{R}}^n(\cdot)\}_{n\in\mathbb{N}}$ is tight. This is done using the same general approach as that outlined for proving tightness of $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$. To illustrate similarity of proof techniques, we consecutively execute each step for proving tightness of both processes in Sections 5.2 through 5.4 below. Then in Section 6, we complete the proof of Lemma 5.3 by uniquely characterizing the limit points.

There are two specific choices of $\{g_{k,i}^n\}_{i\in\mathbb{N}}$, $1 \le k \le K$ and $n \in \mathbb{N}$, that will be of particular interest for proving tightness of $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$ and Theorem 3.2.

Special Case 1: For $n \in \mathbb{N}$, $1 \le k \le K$, and $i \in \mathbb{N}$, let $g_{k,i}^n = d_{k,i}$. This choice clearly satisfies (AA). Hence, Lemma 5.1 implies (42). Further, for $n \in \mathbb{N}$, $1 \le k \le K$, $i \in \mathbb{N}$, and $t \in [0,\infty)$, let

$$a_{k,i}^n(t) = \left(d_{k,i} - \left(t - t_{k,Z_k^n(0)+i}^n \right)^+ \right)^+. \tag{46}$$

Then, for $n \in \mathbb{N}$ and $t \in [0,\infty)$, set

$$\mathcal{A}_k^n(t) = \sum_{i=1}^{E_k^n(t)} \delta^+_{a^n_{k,i}(t)} \qquad \text{and} \qquad \bar{\mathcal{A}}_k^n(t) = \frac{1}{n}\mathcal{A}_k^n(t).$$

By Lemma 5.3, as $n \to \infty$,

$$\bar{\mathcal{A}}^n(\cdot) \Rightarrow \mathcal{R}^*(\cdot). \tag{47}$$

Special Case 2: For $n \in \mathbb{N}$, $1 \le k \le K$, and $i \in \mathbb{N}$, let $g^n_{k,i} = d_{k,i} + v^n_{k,i}$. This choice satisfies (AA) as well. For $n \in \mathbb{N}$, $1 \le k \le K$, $i \in \mathbb{N}$, and $t \in [0,\infty)$, let

$$v^n_{k,i}(t) = \left(d_{k,i} + v^n_{k,i} - \left(t - t^n_{k,Z^n_k(0)+i}\right)^+\right)^+. \tag{48}$$

Then, for $n \in \mathbb{N}$ and $t \in [0,\infty)$, set

$$\mathcal{V}_k^n(t) = \sum_{i=1}^{E_k^n(t)} \delta^+_{v^n_{k,i}(t)} \qquad \text{and} \qquad \bar{\mathcal{V}}_k^n(t) = \frac{1}{n}\mathcal{V}_k^n(t).$$

By Lemma 5.3, as $n \to \infty$,

$$\bar{\mathcal{V}}^n(\cdot) \Rightarrow \mathcal{R}^*(\cdot). \tag{49}$$

Even though the steps for proving tightness of $\{\bar{\mathcal{R}}^n(\cdot)\}_{n\in\mathbb{N}}$ and $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$ are executed consecutively, it is worth noting that Lemma 5.3 is actually used to prove tightness of $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$ through the application of (47) and (49).

5.2. *Compact Containment.* In this section, we demonstrate the compact containment properties needed to prove tightness of $\{\bar{\mathcal{R}}^n(\cdot)\}_{n\in\mathbb{N}}$ and of $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$. In both cases, we will utilize [12, Lemma 15.7.5]. The application turns out to be simpler for $\{\bar{\mathcal{R}}^n(\cdot)\}_{n\in\mathbb{N}}$ since these are measures in $\mathbf{M}_1^K$.

Compact Containment in $\mathbf{M}_1^K$. Let $K_m = [0,m]$ and let K_m^c denote its complement. Then $\{K_m\}_{m\in\mathbb{N}}$ forms a sequence of sets that increases to $\mathbb{R}_+$. By [12, Lemma 15.7.5], $\mathbf{K}' \subset \mathbf{M}_1^K$ is relatively compact if and only if there exists a positive constant $\check{z}$ and a sequence of positive constants $\{b_m\}_{m\in\mathbb{N}}$ tending to zero such that for each $\zeta \in \mathbf{K}'$

$$\max_{1\le k\le K} \zeta_k(\mathbb{R}_+) \le \check{z} \qquad \text{and} \qquad \max_{1\le k\le K} \zeta_k(K_m^c) \le b_m, \ \forall\, m \in \mathbb{N}.$$

Note that for $1 \le k \le K$ and $m \in \mathbb{N}$, $\zeta_k(K_m^c) \le \langle \chi, \zeta_k \rangle / m$. Hence it suffices to show that there exist positive constants $\check{z}$ and $\check{w}$ such that for each $\zeta \in \mathbf{K}'$

$$\max_{1 \le k \le K} \zeta_k(\mathbb{R}_+) \le \check{z} \qquad \text{and} \qquad \max_{1 \le k \le K} \langle \chi, \zeta_k \rangle \le \check{w}. \tag{50}$$

We will verify (50) to prove Lemma 5.4.

LEMMA 5.4. *Suppose that (AA) holds. Let $T, \eta > 0$. There exists a compact set $\mathbf{K} \subset \mathbf{M}_1^K$ such that*

$$\liminf_{n \to \infty} \mathbb{P}(\bar{\mathcal{R}}^n(t) \in \mathbf{K} \text{ for all } t \in [0, T]) \ge 1 - \eta.$$

PROOF. Fix $T, \eta > 0$. For all $1 \le k \le K$ and $t \in [0, T]$,

$$\bar{\mathcal{R}}_k^n(t)(\mathbb{R}_+) \le \bar{E}_k^n(T) \qquad \text{and} \qquad \left\langle \chi, \bar{\mathcal{R}}_k^n(t) \right\rangle \le \left\langle \chi, \bar{\mathcal{G}}_k^n(T) \right\rangle. \tag{51}$$

Define

$$\Omega_1^n = \left\{ \max_{1 \le k \le K} \bar{E}_k^n(T) \le 2\lambda_+ T \right\} \quad \text{and} \quad \Omega_2^n = \left\{ \max_{1 \le k \le K} \left\langle \chi, \bar{\mathcal{G}}_k^n(T) \right\rangle \le 2 g_+ T \right\}.$$

Set $\Omega_0^n = \Omega_1^n \cap \Omega_2^n$. By (45) and Lemma 5.1,

$$\liminf_{n \to \infty} \mathbb{P}\left(\Omega_0^n\right) \ge 1 - \eta. \tag{52}$$

The result follows by combining (50), (51) and (52). □

Compact Containment in $\mathbf{M}_2^K$. For each $m \in \mathbb{N}$, let $K_m = [0, m] \times [0, m]$ and let K_m^c denote its complement. Then $\{K_m\}_{m \in \mathbb{N}}$ forms a sequence of sets that increases to $\mathbb{R}_+^2$. By [12, Lemma 15.7.5], $\mathbf{K}' \subset \mathbf{M}_2^K$ is relatively compact if and only if there exists a positive constant $\check{z}$ and a sequence of positive constants $\{b_m\}_{m \in \mathbb{N}}$ tending to zero as m tends to infinity such that for each $\zeta \in \mathbf{K}'$

$$\max_{1 \le k \le K} \zeta_k(\mathbb{R}_+^2) \le \check{z} \qquad \text{and} \qquad \max_{1 \le k \le K} \zeta_k(K_m^c) \le b_m, \ \forall\, m \in \mathbb{N}. \tag{53}$$

We will verify (53) to prove Lemma 5.5.

LEMMA 5.5. *Let $T, \eta > 0$. There exists a compact set $\mathbf{K} \subset \mathbf{M}_2^K$ such that*

$$\liminf_{n \to \infty} \mathbb{P}(\bar{\mathcal{Z}}^n(t) \in \mathbf{K} \text{ for all } t \in [0, T]) \ge 1 - \eta.$$

PROOF. Fix $T, \eta > 0$. First we identify a sequence of sets $\{\Omega_0^n\}_{n\in\mathbb{N}}$ for which we will show that for each $n \in \mathbb{N}$ on Ω_0^n, $\bar{\mathcal{Z}}^n(t)$ remains in a particular relatively compact set for all $t \in [0, T]$. By (27), there exists a compact set $\mathbf{K}_0$ such that

$$\liminf_{n\to\infty} \mathbb{P}(\bar{\mathcal{Z}}^n(0) \in \mathbf{K}_0) \geq 1 - \frac{\eta}{4}.$$

Since $\mathbf{K}_0$ is compact, there exists a positive constant $\check{z}_0$ and a sequence of positive constants $\{a_m\}_{m\in\mathbb{N}}$ tending to zero as m tends to infinity such that

$$\mathbf{K}_0 \subset \left\{\zeta \in \mathbf{M}_2^K : \max_{1\leq k\leq K} \zeta_k(\mathbb{R}_+^2) \leq \check{z}_0 \text{ and } \max_{1\leq k\leq K} \zeta_k(K_m^c) \leq a_m, \ \forall\, m \in \mathbb{N}\right\}.$$

For $n \in \mathbb{N}$, let

$$\Omega_1^n = \left\{\max_{1\leq k\leq K} \bar{\mathcal{Z}}_k^n(0)(\mathbb{R}_+^2) \leq \check{z}_0 \text{ and } \max_{1\leq k\leq K} \bar{\mathcal{Z}}_k^n(0)(K_m^c) \leq a_m, \ \forall\, m \in \mathbb{N}\right\}.$$

Then

$$\liminf_{n\to\infty} \mathbb{P}\left(\Omega_1^n\right) \geq 1 - \frac{\eta}{4}.$$

For $n \in \mathbb{N}$, let

$$\Omega_2^n = \left\{\max_{1\leq k\leq K} \bar{E}_k^n(T) \leq 2\lambda_+ T\right\}.$$

By (45),

$$\liminf_{n\to\infty} \mathbb{P}\left(\Omega_2^n\right) \geq 1 - \frac{\eta}{4}.$$

Note that, for all $n \in \mathbb{N}$,

$$\sup_{t\in[0,T]} W^n(t) \leq W^n(0) + \frac{1}{n}\sum_{k=1}^K \sum_{i=1}^{E_k^n(T)} v_{k,i}.$$

Then, by (27) and a functional strong law of large numbers, there exists a positive constant $\check{w}$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} W^n(t) \leq \check{w}\right) \geq 1 - \frac{\eta}{4}.$$

For $n \in \mathbb{N}$, let

$$\Omega_3^n = \left\{\sup_{t\in[0,T]} W^n(t) \leq \check{w}\right\}.$$

Notice that for each $n, m \in \mathbb{N}$, $1 \leq k \leq K$, and $t \in [0, T]$,

$$\bar{\mathcal{D}}_k^n(t)((m,\infty)) \leq \bar{\mathcal{D}}_k^n(T)((m,\infty)) \leq \frac{\langle \chi, \bar{\mathcal{D}}_k^n(T)\rangle}{m}. \tag{54}$$

Set $c = 2\lambda_+ g_+ T$. For each $m \in \mathbb{N}$, let $c_m = c/m$. Then (42) and (54) together imply that

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{t\in[0,T]} \bar{\mathcal{D}}_k^n(t)((m,\infty)) \le c_m,\ \forall\ m\in\mathbb{N} \right) \ge 1 - \frac{\eta}{4}.$$

For $n \in \mathbb{N}$, let

$$\Omega_4^n = \left\{ \max_{1\le k\le K} \sup_{t\in[0,T]} \bar{\mathcal{D}}_k^n(t)((m,\infty)) \le c_m,\ \forall\ m\in\mathbb{N} \right\}.$$

Finally, for each $n \in \mathbb{N}$ set

$$\Omega_0^n = \Omega_1^n \cap \Omega_2^n \cap \Omega_3^n \cap \Omega_4^n.$$

It follows that

$$\liminf_{n\to\infty} \mathbb{P}\left(\Omega_0^n\right) \ge 1 - \eta. \tag{55}$$

Next we identify the relatively compact set $\mathbf{K}'$. For this, let $\check{z} = \check{z}_0 + 2\lambda_+ T$ and for $m \in \mathbb{N}$, let

$$b_m = \begin{cases} \check{z}, & 1 \le m \le \check{w}, \\ a_m + c_{m-\check{w}}, & m > \check{w}. \end{cases}$$

Then $\{b_m\}_{m\in\mathbb{N}}$ is a sequence of postive numbers tending to zero as m tends to infinity. Let

$$\mathbf{K}' = \left\{ \zeta \in \mathbf{M}_2^K : \max_{1\le k\le K} \zeta_k(\mathbb{R}_+^2) \le \check{z} \text{ and } \max_{1\le k\le K} \zeta_k(K_m^c) \le b_m, \forall m\in\mathbb{N} \right\}.$$

Then, by (53), $\mathbf{K}'$ is relatively compact.

It suffices to show that for each $n \in \mathbb{N}$, on Ω_0^n, $\bar{\mathcal{Z}}^n(t) \in \mathbf{K}'$ for all $t \in [0,T]$. Fix $n \in \mathbb{N}$. On $\Omega_1^n \cap \Omega_2^n$ we have that, for all $1 \le k \le K$ and $t \in [0,T]$,

$$\bar{\mathcal{Z}}_k^n(t)(\mathbb{R}_+^2) \le \bar{\mathcal{Z}}_k^n(0)(\mathbb{R}_+^2) + \bar{E}_k^n(T) \le \check{z}_0 + 2\lambda_+ T = \check{z}. \tag{56}$$

Next, by (13) under fluid scaling and the fact that for any $x \in \mathbb{R}_+$ and $m \in \mathbb{N}$, $(K_m^c)_x \subseteq K_m^c$, we have, for each $m \in \mathbb{N}$, $1 \le k \le K$, and $t \in [0,T]$,

$$\bar{\mathcal{Z}}_k^n(t)(K_m^c) \quad = \quad \frac{1}{n} \sum_{j=1}^{A_k^n(t)} 1_{(K_m^c)_{t-t_{k,j}^n}}(w_{k,j}^n, p_{k,j}^n)$$

$$\begin{aligned} &\leq \frac{1}{n}\sum_{j=1}^{A_k^n(t)} 1_{K_m^c}(w_{k,j}^n, p_{k,j}^n) \\ &\leq \bar{\mathcal{Z}}_k^n(0)(K_m^c) + \frac{1}{n}\sum_{j=Z_k^n(0)+1}^{A_k^n(t)} \left(1_{\{w_{k,j}^n > m\}} + 1_{\{p_{k,j}^n > m\}}\right). \end{aligned}$$

Recall that on Ω_1^n, $\max_{1\leq k\leq K} \bar{\mathcal{Z}}_k^n(0)(K_m^c) \leq a_m$ for all $m \in \mathbb{N}$. Further, on Ω_3^n, for $m > \breve{w}$, $1_{\{w_{k,j}^n > m\}} = 0$ for each $1 \leq k \leq K$ and $Z_k^n(0)+1 \leq j \leq A_k^n(T)$. Additionally, on Ω_3^n, for each $1 \leq k \leq K$ and $Z_k^n(0)+1 \leq j \leq A_k^n(T)$,

$$p_{k,j}^n \leq d_{k,j-Z_k^n(0)} + W^n(t_{k,j}^n) \leq d_{k,j-Z_k^n(0)} + \breve{w}.$$

Hence, on $\Omega_3^n \cap \Omega_4^n$, for $m > \breve{w}$ and $1 \leq k \leq K$,

$$\frac{1}{n}\sum_{j=Z_k^n(0)+1}^{A_k^n(t)} 1_{\{p_{k,j}^n > m\}} \leq \bar{\mathcal{D}}_k^n(t)((m-\breve{w}, \infty)) \leq c_{m-\breve{w}}.$$

Then on $\Omega_1^n \cap \Omega_3^n \cap \Omega_4^n$ for $m > \breve{w}$,

$$\bar{\mathcal{Z}}_k^n(t)(K_m^c) \leq a_m + c_{m-\breve{w}}. \tag{57}$$

By (56) and (57), it follows that on Ω_0^n, $\bar{\mathcal{Z}}_k^n(t)(K_m^c) \leq b_m$ for all $m \in \mathbb{N}$. This together with (56) implies that on Ω_0^n, $\bar{\mathcal{Z}}^n(t) \in \mathbf{K}'$ for each $t \in [0, T]$. □

5.3. *Asymptotic Regularity.* This section contains results that are preparatory for proving the oscillation bounds. For the sequences of fluid scaled residual deadline related processes and state descriptors, the sudden arrival or departure of a large amount of mass may result in a large oscillation. The focus here is showing that it is very unlikely that large oscillations take place due to departing mass.

Asymptotic Regularity in $\mathbf{M}_1^K$. Consider the sequence $\{\bar{\mathcal{R}}^n(\cdot)\}_{n\in\mathbb{N}}$ of fluid scaled residual deadline related processes. Given $x \in \mathbb{R}_+$ and $\kappa > 0$, let

$$I_x^\kappa = \left((x-\kappa)^+, x+\kappa\right).$$

Note that for $t \in [0,\infty)$, $x \in \mathbb{R}_+$, and $\kappa > 0$, the mass in I_x^κ at time t will all depart the system during the time interval $(t+(x-\kappa)^+, t+x+\kappa)$. This time interval is small if κ is small. Hence, in order to avoid an abrupt departure of a large amount of mass, one needs to show that asymptotically such sets contain arbitrarily small mass. This is stated precisely in the following lemma.

LEMMA 5.6. *Suppose that (AA) holds. Let $T, \varepsilon, \eta > 0$. Then there exists $\kappa > 0$ such that*

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{R}}_k^n(t)(I_x^\kappa) \le \varepsilon \right) \ge 1-\eta.$$

PROOF. Fix $T, \varepsilon, \eta > 0$. Set $\kappa = \frac{\varepsilon}{8\lambda_+}$ and $M_t = \lceil t/\kappa \rceil$ for $0 < t \le T$. For each integer $m \ge 3$, $n \in \mathbb{N}$, $1 \le k \le K$, and $t \in [0,T]$,

$$\bar{\mathcal{G}}_k^n(t)(((m-2)\kappa,\infty)) \le \bar{\mathcal{G}}_k^n(T)(((m-2)\kappa,\infty)) \le \frac{\langle \chi, \bar{\mathcal{G}}_k^n(T)\rangle}{(m-2)\kappa}.$$

It follows by Lemma 5.1 that for some $m \ge 3$,

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{t\in[0,T]} \bar{\mathcal{G}}_k^n(t)(((m-2)\kappa,\infty)) \le \varepsilon \right) \ge 1-\frac{\eta}{2}.$$

Fix such an m and for $n \in \mathbb{N}$, let

$$\Omega_1^n = \left\{ \max_{1\le k\le K} \sup_{t\in[0,T]} \bar{\mathcal{G}}_k^n(t)(((m-2)\kappa,\infty)) \le \varepsilon \right\}.$$

Also, for $n \in \mathbb{N}$, define

$$\begin{aligned}\Omega_2^n \quad = \quad & \left\{ \max_{1\le k\le K} \sup_{0\le s\le t\le T} \max_{0\le j\le m+M_T} \left| \bar{\mathcal{G}}_k^n(s,t)((j\kappa,(j+4)\kappa)) \right.\right. \\ & \left.\left. -\mathcal{D}_k^*(s,t)((j\kappa,(j+4)\kappa)) \right| \le \frac{\varepsilon}{2M_T} \right\}.\end{aligned}$$

By Corollary 5.2,

$$\liminf_{n\to\infty} \mathbb{P}(\Omega_2^n) \ge 1-\frac{\eta}{2}.$$

For $n \in \mathbb{N}$, set $\Omega_0^n = \Omega_1^n \cap \Omega_2^n$. Then, we have that

$$\liminf_{n\to\infty} \mathbb{P}(\Omega_0^n) \ge 1-\eta.$$

Fix $n \in \mathbb{N}$, $1 \le k \le K, x \in \mathbb{R}_+$, and $0 < t \le T$. First, note that for any $1 \le i \le E_k^n(t)$, we have $t_{k,i+Z_k^n(0)}^n \le t$, so that $(t - t_{k,i+Z_k^n(0)}^n)^+ = t - t_{k,i+Z_k^n(0)}^n$. It follows that

$$\bar{\mathcal{R}}_k^n(t)(I_x^\kappa) = \frac{1}{n} \sum_{i=1}^{E_k^n(t)} 1_{I_x^\kappa}(g_{k,i}^n(t)) \le \frac{1}{n} \sum_{i=1}^{E_k^n(t)} 1_{I^\kappa_{x+t-t^n_{k,i+Z_k^n(0)}}}(g_{k,i}^n).$$

Then, for $x \geq m\kappa$ and $1 \leq i \leq E_k^n(t)$,

$$\left(x + t - t^n_{k,i+Z_k^n(0)} - \kappa\right)^+ \geq x - \kappa \geq (m-1)\kappa > (m-2)\kappa.$$

Hence, on Ω_1^n, for $x \geq m\kappa$,

$$\bar{\mathcal{R}}_k^n(t)(I_x^\kappa) \leq \frac{1}{n}\sum_{i=1}^{E_k^n(t)} 1_{((m-2)\kappa,\infty)}(g^n_{k,i}) = \bar{\mathcal{G}}_k^n(t)\left(((m-2)\kappa,\infty)\right) \leq \varepsilon.$$

Otherwise, $x < m\kappa$. Then

$$\begin{aligned}\bar{\mathcal{R}}_k^n(t)(I_x^\kappa) \quad \leq \quad & \frac{1}{n}\left(\sum_{j=0}^{M_t-2}\sum_{i=E_k^n(j\kappa)+1}^{E_k^n((j+1)\kappa)} 1_{I^\kappa_{x+t-t^n_{k,i+Z_k^n(0)}}}(g^n_{k,i})\right.\\ & \left. + \sum_{i=E_k^n((M_t-1)\kappa)+1}^{E_k^n(t)} 1_{I^\kappa_{x+t-t^n_{k,i+Z_k^n(0)}}}(g^n_{k,i})\right).\end{aligned}$$

Noting that for $0 \leq j \leq M_t - 1$ and $E_k^n(j\kappa) < i \leq E_k^n((j+1)\kappa \vee t)$ we have that

$$x + t - \kappa - t^n_{k,i+Z_k^n(0)} \geq x + t - (j+2)\kappa \geq \kappa\left(\left\lfloor\frac{x+t}{\kappa}\right\rfloor - j - 2\right)$$

and

$$x + t + \kappa - t^n_{k,i+Z_k^n(0)} \leq x + t - (j-1)\kappa \leq \kappa\left(\left\lfloor\frac{x+t}{\kappa}\right\rfloor - j + 2\right).$$

Then it follows that

$$\begin{aligned}\bar{\mathcal{R}}_k^n(t)(I_x^\kappa) \quad \leq \quad & \frac{1}{n}\sum_{j=0}^{M_t-2}\sum_{i=E_k^n(j\kappa)+1}^{E_k^n((j+1)\kappa)} 1_{I^{2\kappa}_{\kappa\left(\lfloor\frac{x+t}{\kappa}\rfloor - j\right)}}(g^n_{k,i})\\ & + \frac{1}{n}\sum_{i=E_k^n((M_t-1)\kappa)+1}^{E_k^n(t)} 1_{I^{2\kappa}_{\kappa\left(\lfloor\frac{x+t}{\kappa}\rfloor - (M_t-1)\right)}}(g^n_{k,i})\\ = \quad & \sum_{j=0}^{M_t-2} \bar{\mathcal{G}}_k^n(j\kappa,(j+1)\kappa)\left(I^{2\kappa}_{\kappa\left(\lfloor\frac{x+t}{\kappa}\rfloor - j\right)}\right)\\ & + \bar{\mathcal{G}}_k^n((M_t-1)\kappa, t)\left(I^{2\kappa}_{\kappa\left(\lfloor\frac{x+t}{\kappa}\rfloor - (M_t-1)\right)}\right).\end{aligned}$$

Since $x < m\kappa$, for all $1 \le j \le M_t - 1$, the left end point of $I^{2\kappa}_{\kappa(\lfloor \frac{x+t}{\kappa} \rfloor - j)}$ satisfies

$$0 \le \left(\kappa\left(\left\lfloor \frac{x+t}{\kappa} \right\rfloor - j\right) - 2\kappa\right)^+ \le x + t < m\kappa + M_t\kappa \le (m + M_T)\kappa.$$

Also note that $0 < t - (M_t - 1)\kappa \le t - (t/\kappa - 1)\kappa = \kappa$. Then, on Ω_2^n, since $x < m\kappa$,

$$\begin{aligned}
\bar{\mathcal{R}}_k^n(t)(I_x^\kappa) &\le \sum_{j=0}^{M_t-2} \mathcal{D}_k^*(j\kappa, (j+1)\kappa)\left(I^{2\kappa}_{\kappa(\lfloor \frac{x+t}{\kappa} \rfloor - j)}\right) \\
&\quad + \mathcal{D}_k^*((M_t - 1)\kappa, t)\left(I^{2\kappa}_{\kappa(\lfloor \frac{x+t}{\kappa} \rfloor - (M_t-1))}\right) + \frac{\varepsilon}{2} \\
&\le \sum_{j=0}^{M_t-1} \lambda_k \kappa \left[F_k\left(\kappa\left(\left\lfloor \frac{x+t}{\kappa} \right\rfloor - j + 2\right)^+\right)\right. \\
&\qquad\qquad \left. - F_k\left(\kappa\left(\left\lfloor \frac{x+t}{\kappa} \right\rfloor - j - 2\right)^+\right)\right] + \frac{\varepsilon}{2} \\
&\le \lambda_k \kappa \sum_{j=-2}^{\infty} \left[F_k((j+4)\kappa) - F_k(j\kappa)\right] + \frac{\varepsilon}{2}.
\end{aligned}$$

Note that $F_k(\cdot)$ is a cumulative distribution function and that each point in $\mathbb{R}_+$ is included in at most four intervals of the form $(j\kappa, (j+4)\kappa]$, $j = -2, -1, 0, 1, 2, \ldots$. Thus, on Ω_2^n, since $x < m\kappa$,

$$\bar{\mathcal{R}}_k^n(t)(I_x^\kappa) \le 4\lambda_k\kappa + \frac{\varepsilon}{2} \le \varepsilon.$$

Since, $n \in \mathbb{N}$, $1 \le k \le K$, $x \in \mathbb{R}_+$, and $t \in [0, T]$ were chosen arbitrarily, this concludes the proof. $\square$

Asymptotic Regularity in $\mathbf{M}_2^K$*.* We need to prove an analog of Lemma 5.6 for $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$. In particular, we wish to prove the following prelimit version of Lemma 4.1.

LEMMA 5.7. *Let $T, \varepsilon, \eta > 0$. Then there exists $\kappa > 0$ such that*

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{x\in\mathbb{R}_+^2} \sup_{t\in[0,T]} \bar{\mathcal{Z}}_k^n(t)(C_x^\kappa) \le \varepsilon\right) \ge 1 - \eta.$$

Before proving Lemma 5.7, we verify the following regularity result for the initial state, which is the stochastic analog of (24).

LEMMA 5.8. *Let $\varepsilon, \eta > 0$. Then there exists $\kappa > 0$ such that*

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{x\in\mathbb{R}_+^2} \bar{\mathcal{Z}}_k^n(0)(C_x^\kappa) \le \varepsilon\right) \ge 1-\eta.$$

PROOF. Fix $\varepsilon, \eta > 0$. Given $i = 1, 2$, recall definition (25) of the projection mapping $\pi_i : \mathbf{M}_2 \to \mathbf{M}_1$. We apply the argument given in [8, Pages 835–836] for measures in $\mathbf{M}_1$ to the projection mappings applied to $\bar{\mathcal{Z}}_+^n(0)$, $n \in \mathbb{N}$, to verify that there exists $\kappa > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{i=1,2} \sup_{x\in\mathbb{R}_+} \left\langle 1_{[(x-\kappa)^+, x+\kappa]}, \pi_i\left(\bar{\mathcal{Z}}_+^n(0)\right)\right\rangle < \frac{\varepsilon}{2}\right) \ge 1-\eta. \tag{58}$$

The desired result follows from (58) since for all $n \in \mathbb{N}$, $1 \le k \le K$, $x = (x_1, x_2) \in \mathbb{R}_+^2$ and $\kappa > 0$,

$$\bar{\mathcal{Z}}_k^n(0)(C_x^\kappa) \le \left\langle 1_{[(x_1-\kappa)^+, x_1+\kappa]}, \pi_1\left(\bar{\mathcal{Z}}_+^n(0)\right)\right\rangle + \left\langle 1_{[(x_2-\kappa)^+, x_2+\kappa]}, \pi_2\left(\bar{\mathcal{Z}}_+^n(0)\right)\right\rangle.$$

In order to verify (58), we must verify that suitable K-dimensional analogs of [8, (3.19)–(3.22)] hold. For this, for $\zeta \in \mathbf{M}_2^K$ and $i = 1, 2$ we adopt the shorthand notation

$$\pi_i(\zeta) = (\pi_i(\zeta_1), \ldots, \pi_i(\zeta_K)) \text{ and } \langle \chi, \pi_i(\zeta)\rangle = (\langle \chi, \pi_i(\zeta_1)\rangle, \ldots, \langle \chi, \pi_i(\zeta_K)\rangle).$$

Note that for $\nu \in \mathbf{M}_2$ and $i = 1, 2$, $\langle \chi, \pi_i(\nu)\rangle = \langle \chi \circ p_i, \nu\rangle = \langle p_i, \nu\rangle$. Then, by (27), as $n \to \infty$,

$$\begin{aligned}&\left(\pi_1(\bar{\mathcal{Z}}^n(0)), \pi_2(\bar{\mathcal{Z}}^n(0)), \left\langle \chi, \pi_1(\bar{\mathcal{Z}}^n(0))\right\rangle, \left\langle \chi, \pi_2(\bar{\mathcal{Z}}^n(0))\right\rangle\right)\\ &\quad\Rightarrow (\pi_1(\mathcal{Z}_0^*), \pi_1(\mathcal{Z}_0^*), \langle \chi, \pi_1(\mathcal{Z}_0^*)\rangle, \langle \chi, \pi_1(\mathcal{Z}_0^*)\rangle),\end{aligned}$$

which is the K-dimensional analog of [8, (3.19)]. Using (A.3) we obtain the following K-dimensional analog of [8, (3.20)]:

$$\max_{1\le k\le K} \max_{i=1,2} \mathbb{E}\left[\left\langle 1, \pi_i(\mathcal{Z}_{0,k}^*)\right\rangle\right] \le \mathbb{E}\left[\mathcal{Z}_{0,+}^*(\mathbb{R}_+^2)\right] < \infty.$$

Using (A.2) we obtain the following K-dimensional analog of [8, (3.21)]:

$$\max_{1\le k\le K} \max_{i=1,2} \mathbb{E}\left[\left\langle \chi, \pi_i(\mathcal{Z}_{0,k}^*)\right\rangle\right] \le \mathbb{E}\left[\left\langle p_1 + p_2, \mathcal{Z}_{0,+}^*\right\rangle\right] < \infty.$$

Using (A.1) (and in particular (I.1)) gives the following K-dimensional analog of [8, (3.22)]:

$$\mathbb{P}\left(\max_{1\le k\le K} \max_{i=1,2} \sup_{x\in\mathbb{R}_+} \left\langle 1_{\{x\}}, \pi_i(\mathcal{Z}_{0,k}^*)\right\rangle = 0\right) = 1. \qquad \square$$

Before moving on to prove Lemma 5.7, we obtain an almost sure upper bound on the mass in $C^{\kappa}_{(x,y)}$ in an arbitrary coordinate of the nth system at time t for each $\kappa > 0$, $x, y \in \mathbb{R}_+$, and $t \in [0, \infty)$. To this end, let $n \in \mathbb{N}$, $1 \le k \le K$, $t \in [0, \infty)$, $x, y \in \mathbb{R}_+$, and $\kappa > 0$. The nth system analog of (15) for the set $C^{\kappa}_{(x,y)}$ is

$$\begin{aligned} \mathcal{Z}^n_k(t)(C^{\kappa}_{(x,y)}) &\le \mathcal{Z}^n_k(0)\left(\left(C^{\kappa}_{(x,y)}\right)_t\right) + \sum_{j=Z^n_k(0)+1}^{A^n_k(t)} 1_{\left(C^{\kappa}_{(x,y)}\right)_{t-t^n_{k,j}}}(w^n_{k,j}, p^n_{k,j}) \\ &= \mathcal{Z}^n_k(0)\left(\left(C^{\kappa}_{(x,y)}\right)_t\right) + \sum_{j=Z^n_k(0)+1}^{A^n_k(t)} 1_{C^{\kappa}_{(x,y)}}(w^n_{k,j}(t), p^n_{k,j}(t)). \end{aligned}$$

But then, since the (x, y)-shift of a set followed by the κ-enlargement contains the κ-enlargement followed by the (x, y)-shift,

$$\mathcal{Z}^n_k(t)(C^{\kappa}_{(x,y)}) \le \mathcal{Z}^n_k(0)\left(C^{\kappa}_{(x+t,y+t)}\right) + \sum_{j=Z^n_k(0)+1}^{A^n_k(t)} 1_{C^{\kappa}_{(x,y)}}(w^n_{k,j}(t), p^n_{k,j}(t)). \tag{59}$$

We simplify the summation term by focusing on each coordinate separately. For this, we classify jobs by those whose residual patience time causes the associated unit atom to lie in a certain horizontal band and those whose residual virtual sojourn time causes the associated unit atom to lie in a certain vertical band. Then, for each $n \in \mathbb{N}$, $1 \le k \le K$, $t \in [0, \infty)$, $x, y \in \mathbb{R}_+$, and $\kappa > 0$,

$$\begin{aligned} \mathcal{Z}^n_k(t)(C^{\kappa}_{(x,y)}) &\le \mathcal{Z}^n_k(0)(C^{\kappa}_{(x+t,y+t)}) + \sum_{j=Z^n_k(0)+1}^{A^n_k(t)} 1_{I^{\kappa}_x}(w^n_{k,j}(t)) \\ &\quad + \sum_{j=Z^n_k(0)+1}^{A^n_k(t)} 1_{I^{\kappa}_y}(p^n_{k,j}(t)). \end{aligned} \tag{60}$$

We are prepared to prove Lemma 5.7. The proof given below can be regarded as a stochastic version of the proof of Lemma 4.1.

PROOF OF LEMMA 5.7. Fix ε, $\eta > 0$ and $T > \frac{\varepsilon}{24\lambda_+}$. We begin by defining a sequence $\{\Omega^n_0\}_{n\in\mathbb{N}}$ of events on which the prelimit processes satisfy properties analogous to those exhibited by fluid model solutions and used in the proof of Lemma 4.1. By Lemma 5.8, there exists $\kappa_0 > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\max_{1\le k\le K} \sup_{x,y\in\mathbb{R}_+} \bar{\mathcal{Z}}^n_k(0)\left(C^{\kappa_0}_{(x,y)}\right) \le \frac{\varepsilon}{4}\right) \ge 1 - \frac{\eta}{6}.$$

For $n \in \mathbb{N}$, let

$$\Omega_1^n = \left\{ \max_{1 \le k \le K} \sup_{x,y \in \mathbb{R}_+} \bar{\mathcal{Z}}_k^n(0)\left(C_{(x,y)}^{\kappa_0}\right) \le \frac{\varepsilon}{4} \right\}.$$

By (47) and (49), there exists a $\kappa_1 > 0$ such that

$$\liminf_{n \to \infty} \mathbb{P}\left(\max_{1 \le k \le K} \sup_{y \in \mathbb{R}_+} \sup_{t \in [0,T]} \bar{\mathcal{A}}_k^n(t)(I_y^{\kappa_1}) \le \frac{\varepsilon}{4} \right) \ge 1 - \frac{\eta}{6}$$

and

$$\liminf_{n \to \infty} \mathbb{P}\left(\max_{1 \le k \le K} \sup_{y \in \mathbb{R}_+} \sup_{t \in [0,T]} \bar{\mathcal{V}}_k^n(t)(I_y^{\kappa_1}) \le \frac{\varepsilon}{4} \right) \ge 1 - \frac{\eta}{6}.$$

For $n \in \mathbb{N}$, let

$$\Omega_2^n = \left\{ \max_{1 \le k \le K} \sup_{y \in \mathbb{R}_+} \sup_{t \in [0,T]} \bar{\mathcal{A}}_k^n(t)(I_y^{\kappa_1}) \le \frac{\varepsilon}{4} \right\}$$

and

$$\Omega_3^n = \left\{ \max_{1 \le k \le K} \sup_{y \in \mathbb{R}_+} \sup_{t \in [0,T]} \bar{\mathcal{V}}_k^n(t)(I_y^{\kappa_1}) \le \frac{\varepsilon}{4} \right\}.$$

For $n \in \mathbb{N}$, let

$$\Omega_4^n \quad = \quad \left\{ \max_{1 \le k \le K} \sup_{0 \le s \le t \le T} \bar{A}_k^n(t) - \bar{A}_k^n(s) \le 2\lambda_+(t-s) \right\}.$$

Since for all $n \in \mathbb{N}$, $1 \le k \le K$, and $0 \le s \le t \le T$, $\bar{A}_k^n(t) - \bar{A}_k^n(s) = \bar{E}_k^n(t) - \bar{E}_k^n(s)$, (45) implies that

$$\liminf_{n \to \infty} \mathbb{P}(\Omega_4^n) \ge 1 - \frac{\eta}{6}.$$

Next we identify positive constants δ and M, analogous to the constants δ and M defined in the proof of Lemma 4.1. Let $\delta = \frac{\varepsilon}{24\lambda_+}$. Then $\delta < T$. In order to define M, there are two cases to consider, based on the nature of the abandonment distributions.

Case 1: First suppose that $d_{\max} < \infty$. Let $w_{\max}(\cdot)$ denote the maximal workload fluid model solution, i.e., the workload fluid model solution such that $w_{\max}(0) = d_{\max}$. By the relative ordering property of workload fluid model solutions, it follows that for any workload fluid model solution $w(\cdot)$,

$w(t) \leq w_{\max}(t)$ for all $t \in [0, \infty)$. In particular, for any workload fluid model solution $w(\cdot)$ for all $t \geq \delta$,

$$w(t) \leq w_{\max}(\delta). \tag{61}$$

Set $M_1 = (w_{\max}(\delta) + d_{\max})/2$. By monotonicity properties of workload fluid model solutions, $w_u < w_{\max}(\delta) < M_1 < d_{\max}$. It follows from (27), (A.1), Theorem 2.2, and (61) that

$$\mathbb{P}\left(\sup_{t\in[\delta,T]} W^*(t) < M_1\right) = 1.$$

Case 2: Next consider the case $d_{\max} = \infty$. By (27) and (A.1), there exists $M_2 > w_u$ such that

$$\mathbb{P}\left(W^*(0) < M_2\right) \geq 1 - \frac{\eta}{6}.$$

Since $M_2 > w_u$, (27), (A.1), Theorem 2.2 and monotonicity properties of workload fluid model solutions imply that

$$\mathbb{P}\left(\sup_{t\in[0,T]} W^*(t) < M_2\right) = \mathbb{P}\left(W^*(0) < M_2\right) \geq 1 - \frac{\eta}{6}.$$

Set

$$M = \begin{cases} M_1, & \text{if } d_{\max} < \infty, \\ M_2, & \text{if } d_{\max} = \infty. \end{cases}$$

Now we proceed to bound the prelimt processes by M with probability asymptotically close to one. For $n \in \mathbb{N}$, let

$$\Omega_5^n = \left\{\sup_{t\in[\delta,T]} W^n(t) < M\right\}.$$

Note that by (27), (A.1), Theorem 2.2 and that fact that workload fluid model solutions are continuous, the convergence in distribution in (21) takes place with respect to the topology of uniform convergence on compact sets. Furthermore, the set $\{f \in \mathbf{D}([0,\infty), \mathbb{R}_+) : \sup_{t\in[\delta,T]} f(t) < M\}$ is open with respect to this topology. Then by the Portmanteau theorem

$$\liminf_{n\to\infty} \mathbb{P}\left(\Omega_5^n\right) \geq 1 - \frac{\eta}{6}.$$

Set

$$c = \sum_{k=1}^K \rho_k G_k(M) \qquad \text{and} \qquad \kappa = \min\left(\kappa_0, \kappa_1, \frac{\varepsilon}{24\lambda_+}, \frac{\varepsilon c}{72\lambda_+}\right).$$

Note that $c > 0$ since $M < d_{\max}$, and so $\kappa > 0$. Finally, for $n \in \mathbb{N}$, let

$$\Omega_6^n = \left\{ \sup_{t\in[0,T]} |X^n(t)| \le \frac{\kappa}{2} \right\}.$$

Then, by (20),

$$\liminf_{n\to\infty} \mathbb{P}\left(\Omega_6^n\right) \ge 1 - \frac{\eta}{6}.$$

For $n \in \mathbb{N}$, set

$$\Omega_0^n = \Omega_1^n \cap \Omega_2^n \cap \Omega_3^n \cap \Omega_4^n \cap \Omega_5^n \cap \Omega_6^n.$$

Then

$$\liminf_{n\to\infty} \mathbb{P}(\Omega_0^n) \ge 1 - \eta. \tag{62}$$

Fix $n \in \mathbb{N}$ such that $1/n \le \varepsilon/12$, $1 \le k \le K$, $x, y \in \mathbb{R}_+$, and $t \in [0, T]$. The fluid scaled analog of (60) is

$$\begin{aligned}\bar{\mathcal{Z}}_k^n(t)\left(C^{\kappa}_{(x,y)}\right) \quad &\le \quad \bar{\mathcal{Z}}_k^n(0)\left(C^{\kappa}_{(x+t,y+t)}\right) + \frac{1}{n}\sum_{j=Z_k^n(0)+1}^{A_k^n(t)} 1_{I_x^\kappa}\left(w_{k,j}^n(t)\right) \\ &\quad + \frac{1}{n}\sum_{j=Z_k^n(0)+1}^{A_k^n(t)} 1_{I_y^\kappa}\left(p_{k,j}^n(t)\right).\end{aligned} \tag{63}$$

We will show that on Ω_0^n, the right hand side of (63) is less than or equal to ε. Since $\kappa \le \kappa_0$, we have $C^{\kappa}_{(x+t,y+t)} \subseteq C^{\kappa_0}_{(x+t,y+t)}$. Then, on Ω_1^n,

$$\bar{\mathcal{Z}}_k^n(0)\left(C^{\kappa}_{(x+t,y+t)}\right) \le \frac{\varepsilon}{4}. \tag{64}$$

Also, for all $Z_k^n(0)+1 \le j \le A_k^n(t)$, the quantity $p_{k,j}^n(t)$ is either $a^n_{k,j-Z_k^n(0)}(t)$ of $v^n_{k,j-Z_k^n(0)}(t)$ (recall (46) and (48)). It follows that, on $\Omega_2^n \cap \Omega_3^n$,

$$\frac{1}{n}\sum_{j=Z_k^n(0)+1}^{A_k^n(t)} 1_{I_y^\kappa}\left(p_{k,j}^n(t)\right) \le \bar{\mathcal{A}}_k^n(t)\left(I_y^\kappa\right) + \bar{\mathcal{V}}_k^n(t)\left(I_y^\kappa\right) \le \frac{\varepsilon}{2}. \tag{65}$$

Combining (63), (64), and (65), we see that on Ω_0^n,

$$\bar{\mathcal{Z}}_k^n(t)\left(C^{\kappa}_{(x,y)}\right) \le \frac{3\varepsilon}{4} + \frac{1}{n}\sum_{j=Z_k^n(0)+1}^{A_k^n(t)} 1_{I_x^\kappa}\left(w_{k,j}^n(t)\right). \tag{66}$$

Finally, we bound the second term on the right side of (66). To this end, define the prelimit version $\tau^n(\cdot)$ of $\tau(\cdot)$ as follows:

$$\tau^n(s) = \inf\{u \geq 0 : W^n(u) + u \geq s\}, \qquad s \in [0, \infty).$$

Notice that for all $s \in [0, \infty)$, $W^n(s) + s \geq s$, so that $\tau^n(s) \in [0, s]$. During busy periods, $W^n(\cdot) + \iota(\cdot)$ jumps up exactly when jobs arrive that contribute to the workload, and remains constant otherwise. During idle periods, $W^n(\cdot) + \iota(\cdot)$ increases at rate one. In particular, $W^n(\cdot) + \iota(\cdot)$ is right continuous and nondecreasing. Then,

$$W^n(\tau^n(s)) + \tau^n(s) \geq s, \qquad \text{for } s \in [0, \infty), \tag{67}$$

$$W^n(\tau^n(s)-) + \tau^n(s) \leq s, \qquad \text{for } s \in (W^n(0), \infty). \tag{68}$$

The function $\tau^n(\cdot)$ will be used to determine which arrivals may contribute to the sum in (66). First we show that

$$\frac{1}{n} \sum_{j=Z_k^n(0)+1}^{A_k^n(t)} 1_{I_x^\kappa}\left(w_{k,j}^n(t)\right) \leq \bar{A}_k^n(\tau^n(x+\kappa+t) \wedge t) - \bar{A}_k^n\left(\tau^n((x-\kappa)^+ + t) \wedge t\right) + \frac{1}{n}. \tag{69}$$

To see this, let $Z_k^n(0) + 1 \leq j \leq A_k^n(t)$. Then $t_{k,j}^n \leq t$ and

$$(x-\kappa)^+ < w_{k,j}^n(t) < x+\kappa \qquad \Leftrightarrow \qquad (x-\kappa)^+ + t < w_{k,j}^n + t_{k,j}^n < x+\kappa+t.$$

But $w_{k,j}^n = W^n(t_{k,j}^n)$. Then, since $W^n(\cdot) + \iota(\cdot)$ is nondecreasing, $t_{k,j}^n \leq t$ and $w_{k,j}^n(t) \in I_x^\kappa$ imply that

$$\tau^n\left((x-\kappa)^+ + t\right) \wedge t \leq t_{k,j}^n \leq \tau^n(x+\kappa+t) \wedge t.$$

Therefore, (69) holds.

Next, we proceed to bound the right hand side of (69) on Ω_0^n. By (66), (69), the definition of Ω_4^n, and the fact that n is such that $1/n \leq \varepsilon/12$, it suffices to show that on Ω_0^n,

$$2\lambda_+\left(\tau^n(x+\kappa+t) \wedge t - \tau^n((x-\kappa)^+ + t) \wedge t\right) \leq \frac{\varepsilon}{6}. \tag{70}$$

If $\tau^n((x-\kappa)^+ + t) \geq t$, then the left side of (70) is zero, and so (70) holds. Henceforth, we assume that $\tau^n((x-\kappa)^+ + t) < t$. If $\tau^n(x+\kappa+t) \wedge t \leq \delta$, then (70) holds since $\delta = \varepsilon/24\lambda_+$, which implies that the left hand side of (70) is no larger than $\varepsilon/12$. Otherwise, $\delta < \tau^n(x+\kappa+t) \wedge t$ (so that

$\tau^n(x+\kappa+t)>0$ and then $x+\kappa+t>W^n(0)$). First consider the case where $\delta\le\tau^n((x-\kappa)^++t)$. Then, by (67), (68), and the nondecreasing nature of $W^n(\cdot)+\iota(\cdot)$,

$$\begin{aligned}2\kappa &= x+\kappa+t-(x-\kappa+t)\ge x+\kappa+t-((x-\kappa)^++t)\\ &\ge W^n(\tau^n(x+\kappa+t)-)+\tau^n(x+\kappa+t)\\ &\quad -W^n(\tau^n((x-\kappa)^++t))-\tau^n((x-\kappa)^++t)\\ &\ge W^n(\tau^n(x+\kappa+t)\wedge t-)+\tau^n(x+\kappa+t)\wedge t\\ &\quad -W^n(\tau^n((x-\kappa)^++t))-\tau^n((x-\kappa)^++t).\end{aligned}$$

Using (19) and the nondecreasing nature of the idle time process, we obtain

$$\begin{aligned}2\kappa &\ge X^n(\tau^n(x+\kappa+t)\wedge t-)-X^n(\tau^n((x-\kappa)^++t))\\ &\quad +\sum_{k=1}^K\rho_k\int_{\tau^n((x-\kappa)^++t)}^{\tau^n(x+\kappa+t)\wedge t}G_k(W^n(u))\mathrm{d}u.\end{aligned}$$

Since $\delta\le\tau^n((x-\kappa)^++t)$, on $\Omega_5^n\cap\Omega_6^n$,

$$2\kappa\ge-\kappa+c\left(\tau^n(x+\kappa+t)\wedge t-\tau^n((x-\kappa)^++t)\right).$$

By definition of κ, $3\kappa/c\le\varepsilon/24\lambda_+$, which implies that the left side of (70) is no larger than $\varepsilon/12$ on $\Omega_5^n\cap\Omega_6^n$, and so (70) holds. The last case to consider is the case $\tau^n((x-\kappa)^++t)<\delta<\tau^n(x+\kappa+t)\wedge t$. Then, by definition of δ, on Ω_4^n,

$$\begin{aligned}&2\lambda_+\left(\tau^n(x+\kappa+t)\wedge t-\tau^n((x-\kappa)^++t)\wedge t\right)\\ &\quad\le\frac{\varepsilon}{12}+2\lambda_+\left(\tau^n(x+\kappa+t)\wedge t-\delta\right).\end{aligned}\tag{71}$$

By (67), (68), and monotonicity of $W^n(\cdot)+\iota(\cdot)$,

$$\begin{aligned}x-\kappa+t &\le (x-\kappa)^++t\\ &\le W^n(\tau^n((x-\kappa)^++t))+\tau^n((x-\kappa)^++t)\\ &\le W^n(\delta)+\delta\\ &\le W^n(\tau^n(x+\kappa+t)\wedge t-)+\tau^n(x+\kappa+t)\wedge t\le x+\kappa+t.\end{aligned}$$

This together with (19) and the nondecreasing nature of the idle time process implies that, on $\Omega_5^n\cap\Omega_6^n$,

$$\begin{aligned}2\kappa &\ge W^n(\tau^n(x+\kappa+t)\wedge t-)+\tau^n(x+\kappa+t)\wedge t-W^n(\delta)-\delta\\ &\ge X^n(\tau^n(x+\kappa+t)\wedge t-)-X^n(\delta)\end{aligned}$$

$$
\begin{aligned}
&\quad + \sum_{k=1}^{K} \rho_k \int_{\delta}^{\tau^n(x+\kappa+t)\wedge t} G_k(W^n(u))\mathrm{d}u \\
&\geq \quad -\kappa + c\left(\tau^n(x+\kappa+t)\wedge t - \delta\right).
\end{aligned}
$$

Using the definition of κ and (71) implies (70). □

5.4. *Oscillation Bounds.* This section establishes the second main ingredient for proving tightness of the fluid scaled residual deadline related processes and state descriptors, controlled oscillations (see Lemmas 5.10 and 5.11). Both proofs are similar in spirit. The one for the residual deadline processes is slightly simpler, so it is presented first. For this, recall the definition of $\mathbf{d}_K$ given in (40). Then the modulus of continuity is defined as follows.

DEFINITION 5.9. Let $i = 1, 2$. For each $\zeta(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M}_i^K)$ and each $T > \delta > 0$, define the modulus of continuity on $[0, T]$ by

$$
\mathbf{w}_T(\zeta(\cdot), \delta) = \sup_{t\in[0,T-\delta]} \sup_{h\in[0,\delta]} \mathbf{d}_K[\zeta(t+h), \zeta(t)].
$$

LEMMA 5.10. *Suppose that (AA) holds. For all $T > 0$ and $\varepsilon, \eta \in (0,1)$, there exists $\delta \in (0, T)$ such that*

$$
\liminf_{n\to\infty} \mathbb{P}(\mathbf{w}_T(\bar{\mathcal{R}}^n(\cdot), \delta) \leq \varepsilon) \geq 1 - \eta.
$$

PROOF. Fix $T > 0$ and $\varepsilon, \eta \in (0,1)$. By (41) and Lemma 5.6, there exists a $\kappa \in (0, \varepsilon)$ such that for any $\delta \in (0, T)$, the events

$$
\begin{aligned}
\Omega_1^n &= \left\{ \max_{1\leq k\leq K} \sup_{t\in[0,T]} \bar{\mathcal{R}}_k^n(t)([0,\kappa]) \leq \varepsilon \right\}, \\
\Omega_2^n &= \left\{ \max_{1\leq k\leq K} \sup_{t\in[0,T-\delta]} \sup_{h\in[0,\delta]} \left[\bar{E}_k^n(t+h) - \bar{E}_k^n(t)\right] \leq 2\lambda_+\delta \right\}, \\
\Omega_0^n &= \Omega_1^n \cap \Omega_2^n,
\end{aligned}
$$

satisfy

$$
\liminf_{n\to\infty} \mathbb{P}(\Omega_0^n) \geq 1 - \eta. \tag{72}
$$

Fix such a κ and set $\delta = \min(\kappa, \frac{\varepsilon}{2\lambda_+})$.

Fix $n \in \mathbb{N}$, $1 \leq k \leq K$, and a closed set $B \in \mathcal{B}_1$. Let $t \in [0, T-\delta]$, $h \in (0, \delta]$, and $1 \leq i \leq E_k^n(t)$. If $g_{k,i}^n(t) > h$ or $g_{k,i}^n(t+h) > 0$, then

$$g_{k,i}^n(t) = g_{k,i}^n(t+h) + h.$$

Then, for $h \in (0, \delta]$, since $h \leq \delta \leq \kappa < \varepsilon$,

$$\begin{aligned} \bar{\mathcal{R}}_k^n(t)(B) &\leq \bar{\mathcal{R}}_k^n(t+h)(B^\varepsilon) + \bar{\mathcal{R}}^n(t)([0,\kappa]) \\ \bar{\mathcal{R}}_k^n(t+h)(B) &\leq \bar{\mathcal{R}}_k^n(t)(B^\varepsilon) + \bar{E}_k^n(t+h) - \bar{E}_k^n(t). \end{aligned}$$

Then, on Ω_0^n, for all $t \in [0, T-\delta]$ and $h \in [0, \delta]$,

$$\mathbf{d}(\bar{\mathcal{R}}_k^n(t+h), \bar{\mathcal{R}}_k^n(t)) < \varepsilon.$$

This together with the fact that $n \in \mathbb{N}$, $1 \leq k \leq K$, and the closed set $B \in \mathcal{B}_2$ were arbitrary and (72) implies the result. $\square$

Next we generalize the preceding argument to prove the following analogous lemma for the fluid scaled state descriptors. There are two main distinctions. One is that $\kappa < \varepsilon/2$ since in $h > 0$ time units a unit atom moves along a diagonal path a distance of $\sqrt{2}h$. The other is that (15) has been established.

LEMMA 5.11. *For all $T > 0$ and $\varepsilon, \eta \in (0,1)$, there exists $\delta \in (0, T)$ such that*

$$\liminf_{n\to\infty} \mathbb{P}(\mathbf{w}_T(\bar{\mathcal{Z}}^n(\cdot), \delta) \leq \varepsilon) \geq 1 - \eta.$$

PROOF. Fix $T > 0$ and $\varepsilon, \eta \in (0,1)$. For each $\kappa > 0$, let $\bar{C}^\kappa$ be the closure of C^κ, or

$$\bar{C}^\kappa = [0,\kappa] \times \mathbb{R}_+ \cup \mathbb{R}_+ \times [0,\kappa].$$

By (41) and Lemma 5.7, there exists a $\kappa \in (0, \varepsilon/2)$ such that for any $\delta \in (0, T)$, the events

$$\begin{aligned} \Omega_1^n &= \left\{ \max_{1\leq k\leq K} \sup_{t\in[0,T]} \bar{\mathcal{Z}}_k^n(t)(\bar{C}^\kappa) \leq \varepsilon \right\}, \\ \Omega_2^n &= \left\{ \max_{1\leq k\leq K} \sup_{t\in[0,T-\delta]} \sup_{h\in[0,\delta]} \left[\bar{E}_k^n(t+h) - \bar{E}_k^n(t)\right] \leq 2\lambda_+\delta \right\}, \\ \Omega_0^n &= \Omega_1^n \cap \Omega_2^n, \end{aligned}$$

satisfy

$$\liminf_{n\to\infty} \mathbb{P}(\Omega_0^n) \geq 1 - \eta. \tag{73}$$

Fix such a κ and set $\delta = \min(\kappa, \frac{\varepsilon}{2\lambda_+})$.

We begin by noting two basic facts that will be used in the proof. Firstly, for all $B \in \mathcal{B}_2$ and $h \in [0, \delta]$,

$$B \subseteq (B^\varepsilon)_h \cup \bar{C}^\kappa. \tag{74}$$

To see this, take some $B \in \mathcal{B}_2$, $h \in [0, \delta]$, and $(w, p) \in B \backslash \bar{C}^\kappa$. By the construction of $\bar{C}^\kappa$, we have $w, p > \kappa \geq \delta \geq h$. Because $h \leq \delta \leq \kappa < \varepsilon/2$, it follows that $(w - h, p - h) \in B^\varepsilon$ and $(w, p) \in (B^\varepsilon)_h$. In addition, since $\delta < \varepsilon/2$, we have that for all $B \in \mathcal{B}_2$ and $h \in [0, \delta]$,

$$B_h \subseteq B^\varepsilon. \tag{75}$$

Fix $n \in \mathbb{N}$, $1 \leq k \leq K$, and a closed set $B \in \mathcal{B}_2$. Let $\check{B} = B \setminus C$. Note that $\check{B} \in \mathcal{B}_{2,0}$ and by (11), $\bar{\mathcal{Z}}_k^n(t)(B) = \bar{\mathcal{Z}}_k^n(t)(\check{B})$ for all $t \in [0, T]$. Then, we can then conclude from (11), (15), and (74) that, on Ω_0^n, for all $t \in [0, T - \delta]$ and $h \in [0, \delta]$,

$$\begin{aligned} \bar{\mathcal{Z}}_k^n(t)(B) &= \bar{\mathcal{Z}}_k^n(t)(\check{B}) \\ &\leq \bar{\mathcal{Z}}_k^n(t)((\check{B}^\varepsilon)_h) + \bar{\mathcal{Z}}_k^n(t)(\bar{C}^\kappa) \\ &\leq \bar{\mathcal{Z}}_k^n(t+h)(\check{B}^\varepsilon) + \varepsilon \\ &\leq \bar{\mathcal{Z}}_k^n(t+h)(B^\varepsilon) + \varepsilon. \end{aligned} \tag{76}$$

Also, by (11), (15), (75), and the fact that $\delta < \frac{\varepsilon}{2\lambda_+}$, it is true that on Ω_0^n, for all $t \in [0, T - \delta]$ and $h \in [0, \delta]$,

$$\begin{aligned} \bar{\mathcal{Z}}_k^n(t+h)(B) &= \bar{\mathcal{Z}}_k^n(t+h)(\check{B}) \\ &\leq \bar{\mathcal{Z}}_k^n(t)(\check{B}_h) + \bar{E}_k^n(t+h) - \bar{E}_k^n(t) \\ &\leq \bar{\mathcal{Z}}_k^n(t)(B_h) + \varepsilon \\ &\leq \bar{\mathcal{Z}}_k^n(t)(B^\varepsilon) + \varepsilon. \end{aligned} \tag{77}$$

Because k and B were chosen arbitrarily, (76) and (77) imply that on Ω_0^n

$$\mathbf{w}_T(\bar{\mathcal{Z}}^n(\cdot), \delta) \leq \varepsilon.$$

The result follows from this and (73). ☐

6. Characterization of Limit Points. The main goal of this section is to prove Theorem 3.2. First we prove Lemma 5.3, which is then used in the proof of Theorem 3.2.

PROOF OF LEMMA 5.3. Throughout, we assume that (AA) holds. Together Lemmas 5.4 and 5.10 imply tightness of $\{\bar{\mathcal{R}}^n(\cdot)\}_{n\in\mathbb{N}}$. Let $\mathbb{M}\subset\mathbb{N}$ be a strictly increasing subsequence tending to infinity and $\bar{\mathcal{R}}(\cdot)$ a process such that as $m\to\infty$,

$$\bar{\mathcal{R}}^m(\cdot)\Rightarrow\bar{\mathcal{R}}(\cdot).$$

By Lemma 5.6, $\bar{\mathcal{R}}(t)$ doesn't charge points for all $t\in[0,\infty)$ almost surely. By (41) and Lemma 5.1 and the deterministic nature of those limiting processes, as $m\to\infty$,

$$\left(\bar{\mathcal{R}}^m(\cdot),\bar{E}^m(\cdot),\bar{\mathcal{G}}^m(\cdot),\left\langle\chi,\bar{\mathcal{G}}^m(\cdot)\right\rangle\right)\Rightarrow\left(\bar{\mathcal{R}}(\cdot),E^*(\cdot),\mathcal{D}^*(\cdot),\left\langle\chi,\mathcal{D}^*(\cdot)\right\rangle\right). \tag{78}$$

Using the Skorohod representation we may assume without loss of generally that all random elements are defined on a common probability space $(\Omega,\mathcal{F},\mathbb{P})$ such that the joint convergence in (78) is almost sure. Fix $\omega\in\Omega$ such that $\bar{\mathcal{R}}(t)(\omega)$ doesn't charge points for all $t\in[0,\infty)$ and as $m\to\infty$,

$$\begin{aligned}&\left(\bar{\mathcal{R}}^m(\cdot)(\omega),\bar{E}^m(\cdot)(\omega),\bar{\mathcal{G}}^m(\cdot)(\omega),\left\langle\chi,\bar{\mathcal{G}}^m(\cdot)\right\rangle(\omega)\right)\\&\qquad\to\left(\bar{\mathcal{R}}(\cdot)(\omega),E^*(\cdot),\mathcal{D}^*(\cdot),\left\langle\chi,\mathcal{D}^*(\cdot)\right\rangle\right).\end{aligned} \tag{79}$$

Henceforth, all random variables are evaluated at this ω. It suffices to show that $\bar{\mathcal{R}}(\cdot)=\mathcal{R}^*(\cdot)$. For this, it suffices to show that for all $1\le k\le K$, $t\in[0,\infty)$ and $x\in\mathbb{R}_+$,

$$\bar{\mathcal{R}}_k(t)(x,\infty)=\lambda_k\int_0^t G_k(x+t-s)\mathrm{d}s. \tag{80}$$

To see this, note that for all $1\le k\le K$, $t\in[0,\infty)$, and $x\in\mathbb{R}_+$

$$\begin{aligned}\mathcal{R}_k^*(t)(x,\infty)&=\lambda_k\int_x^\infty\left[G_k(y)-G_k(y+t)\right]\mathrm{d}y\\&=\lambda_k\int_x^{x+t}G_k(y)\mathrm{d}y=\lambda_k\int_0^t G_k(x+t-s)\mathrm{d}s.\end{aligned}$$

Next we verify (80). For this, fixed $1\le k\le K$, $t\in[0,\infty)$, and $x\in\mathbb{R}_+$. Given $L\in\mathbb{N}$, let $\kappa=t/L$ and set

$$t_\ell=\ell\kappa\qquad\text{and}\qquad x_\ell=x+t-t_\ell,\qquad\text{for }\ell=0,\dots,L.$$

Then, given $L\in\mathbb{N}$, $1\le i\le E_k^m(t)$ if and only if there exists $\ell\in\{0,\dots,L-1\}$ such that $t^m_{k,Z_k^m(0)+i}\in(t_\ell,t_{\ell+1}]$. Further, given $L\in\mathbb{N}$, $1\le i\le E_k^m(t)$ and $\ell\in\{0,\dots,L-1\}$ such that $t^m_{k,Z_k^m(0)+i}\in(t_\ell,t_{\ell+1}]$, $x_\ell<g^m_{k,i}$ implies that $x<g^m_{k,i}(t)$. Similarly, given $L\in\mathbb{N}$, $1\le i\le E_k^m(t)$ and $\ell\in\{0,\dots,L-1\}$

such that $t^m_{k,Z^m_k(0)+i} \in (t_\ell, t_{\ell+1}]$, $x < g^m_{k,i}(t)$ implies that $x_{\ell+1} < g^m_{k,i}$. Hence, given $L \in \mathbb{N}$,

$$\sum_{\ell=0}^{L-1} \bar{\mathcal{G}}^m_k (t_\ell, t_{\ell+1}) (x_\ell, \infty) \le \bar{\mathcal{R}}^m_k(t)(x,\infty) \le \sum_{\ell=0}^{L-1} \bar{\mathcal{G}}^m_k (t_\ell, t_{\ell+1}) (x_{\ell+1}, \infty).$$

By (79), given $L \in \mathbb{N}$ and $\varepsilon > 0$ there exists $M \in \mathbb{N}$ such that for all $m \ge M$

$$\begin{aligned} \max_{0\le \ell\le L-1} \left|\bar{\mathcal{G}}^m_k (t_\ell, t_{\ell+1}) (x_\ell, \infty) - \mathcal{D}^*_k (t_\ell, t_{\ell+1}) (x_\ell, \infty)\right| &< \frac{\varepsilon}{2L}, \\ \max_{0\le \ell\le L-1} \left|\bar{\mathcal{G}}^m_k (t_\ell, t_{\ell+1}) (x_{\ell+1}, \infty) - \mathcal{D}^*_k (t_\ell, t_{\ell+1}) (x_{\ell+1}, \infty)\right| &< \frac{\varepsilon}{2L}. \end{aligned}$$

Hence, given $L \in \mathbb{N}$ and $\varepsilon > 0$ there exists $M \in \mathbb{N}$ such that for all $m \ge M$,

$$\sum_{\ell=0}^{L-1} \mathcal{D}^*_k (t_\ell, t_{\ell+1}) (x_\ell, \infty) - \frac{\varepsilon}{2} \le \bar{\mathcal{R}}^m_k(t)(x,\infty) \le \sum_{\ell=0}^{L-1} \mathcal{D}^*_k (t_\ell, t_{\ell+1}) (x_{\ell+1}, \infty) + \frac{\varepsilon}{2}.$$

Given $L \in \mathbb{N}$, we have that

$$\begin{aligned} \sum_{\ell=0}^{L-1} \mathcal{D}^*_k (t_\ell, t_{\ell+1}) (x_\ell, \infty) &= \sum_{\ell=0}^{L-1} \lambda_k \kappa G_k (x_\ell) \\ \sum_{\ell=0}^{L-1} \mathcal{D}^*_k (t_\ell, t_{\ell+1}) (x_{\ell+1}, \infty) &= \sum_{\ell=0}^{L-1} \lambda_k \kappa G_k (x_{\ell+1}). \end{aligned}$$

Respectively these are upper and lower Riemann sums, and since $G_k(\cdot)$ is continuous, they both converge to $\lambda_k \int_0^t G_k(x+t-s)\mathrm{d}s$ as $L \to \infty$. Given $\varepsilon > 0$, let $\hat{L} \in \mathbb{N}$ be such that

$$\left|\sum_{\ell=0}^{\hat{L}-1} \lambda_k \kappa G_k (x_{\ell+1}) - \sum_{\ell=0}^{\hat{L}-1} \lambda_k \kappa G_k (x_\ell)\right| < \frac{\varepsilon}{2}.$$

Hence, given $\varepsilon > 0$, there exists $\hat{M} \in \mathbb{N}$ such that for all $m \ge \hat{M}$,

$$\left|\bar{\mathcal{R}}^m_k(t)(x,\infty) - \lambda_k \int_0^t G_k(x+t-s)\mathrm{d}s\right| < \varepsilon.$$

Thus, (80) holds, as desired. □

Having proved Lemma 5.3, we are now ready to prove Theorem 3.2.

PROOF OF THEOREM 3.2. Together Lemmas 5.5 and 5.11 imply tightness of $\{\bar{\mathcal{Z}}^n(\cdot)\}_{n\in\mathbb{N}}$. Let $\mathbb{M} \subset \mathbb{N}$ be a strictly increasing subsequence tending to infinity and $\mathcal{Z}^*(\cdot)$ a process such that as $m \to \infty$,

$$\bar{\mathcal{Z}}^m(\cdot) \Rightarrow \mathcal{Z}^*(\cdot). \tag{81}$$

Note that by (A.1), $\mathcal{Z}^*(\cdot) \in \mathbf{I}$ almost surely. Hence, in order to prove Theorem 3.2, it suffices to show that $\mathcal{Z}^*(\cdot)$ satisfies (26) almost surely. Indeed, once this is verified, it follows by the uniqueness asserted in Theorem 3.1 that the law of the limit point $\mathcal{Z}^*(\cdot)$ is unique, and so $\bar{\mathcal{Z}}^n(\cdot) \Rightarrow \mathcal{Z}^*(\cdot)$ as $n \to \infty$.

By (27), (A.1), and Theorem 2.2, as $m \to \infty$,

$$W^m(\cdot) \Rightarrow W^*(\cdot), \tag{82}$$

where $W^*(\cdot)$ is almost surely a workload fluid model solution such that $W^*(0)$ is equal in distribution to W_0^*. We would like to argue that this convergence is joint with (81). By (27), (20), (47), (49), and the fact that the limit in (20) and $\mathcal{R}^*(\cdot)$ are deterministic, as $m \to \infty$,

$$\left(\bar{\mathcal{Z}}^m(\cdot), W^m(0), X^m(\cdot), \bar{\mathcal{A}}^m(\cdot), \bar{\mathcal{V}}^m(\cdot)\right) \Rightarrow \left(\mathcal{Z}^*(\cdot), W_0, 0, \mathcal{R}^*(\cdot), \mathcal{R}^*(\cdot)\right).$$

This together with (22), (23), and (82) implies that as $m \to \infty$,

$$\begin{aligned}\left(\bar{\mathcal{Z}}^m(\cdot), W^m(\cdot), X^m(\cdot), \bar{\mathcal{A}}^m(\cdot), \bar{\mathcal{V}}^m(\cdot)\right)& \\ \Rightarrow \left(\mathcal{Z}^*(\cdot), W^*(\cdot), 0, \mathcal{R}^*(\cdot), \mathcal{R}^*(\cdot)\right)&.\end{aligned} \tag{83}$$

Using the Skorohod representation we may assume without loss of generally that all random elements are defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$ such that the joint convergence in (83) is almost sure. By (27), $\mathcal{Z}^*(0)$ satisfies $w_{\mathcal{Z}^*(0)} = W^*(0)$, (A.1), (A.2), and (A.3) almost surely. Furthermore, by Lemma 5.11, $\mathcal{Z}^*(\cdot)$ is continuous almost surely. In addition, by Lemma 5.7,

$$\mathbb{P}\left(\mathcal{Z}_+^*(t)(C_x) = 0 \text{ for all } t \in [0,\infty) \text{ and } x \in \mathbb{R}_+^2\right) = 1. \tag{84}$$

(cf. [9, Lemma 6.2]). Fix $\omega \in \Omega$ such that $W^*(\cdot)(\omega)$ is a workload fluid model solution and $\mathcal{Z}^*(\cdot)(\omega)$ is continuous and satisfies (84), $w_{\mathcal{Z}^*(0)(\omega)} = W^*(0)(\omega)$, (A.1), (A.2), (A.3) and, as $m \to \infty$,

$$\begin{aligned}\left(\bar{\mathcal{Z}}^m(\cdot)(\omega), W^m(\cdot)(\omega), \bar{\mathcal{A}}^m(\cdot)(\omega), \bar{\mathcal{V}}^m(\cdot)(\omega)\right)& \\ \to \left(\mathcal{Z}^*(\cdot)(\omega), W^*(\cdot)(\omega), \mathcal{R}^*(\cdot), \mathcal{R}^*(\cdot)\right)&.\end{aligned}$$

For all $t \in [0,\infty)$ and $m \in \mathbb{M}$, set

$$\begin{aligned} \zeta^m(t) &= \bar{\mathcal{Z}}^m(t)(\omega) \qquad \text{and} \qquad w^m(t) = W^m(t)(\omega), \\ \zeta(t) &= \mathcal{Z}^*(t)(\omega) \qquad \text{and} \qquad w(t) = W^*(t)(\omega). \end{aligned}$$

By (27), $w(0) = w_\vartheta$ where $\vartheta = \zeta(0)$. In order to prove Theorem 3.2, it suffices to show that $\zeta(\cdot)$ is a fluid model solution for the supercritical data (λ, μ, Γ) and initial measure ϑ. In particular, we must show that $\zeta(\cdot)$ satisfies (26). In this regard, recall that $\mathcal{P}$ is a π-system (see (35)). As in the proof of Theorem 3.1 given in Section 4, it is enough to show that $\zeta(\cdot)$ satisfies (26) for all $B \in \mathcal{P}$.

Fix $B \in \mathcal{P}$. Then $B = [a,\infty) \times [c,\infty)$ for some $0 \le a, c < \infty$. Fix $t \in [0,\infty)$. In what follows, all random elements are evaluated at the specific ω fixed in the preceding paragraph. Since $\zeta(\cdot)$ is continuous, $\zeta^n(s) \overset{w}{\to} \zeta(s)$ for all $s \in [0,t]$. By (84), $\zeta_+(0)(C_{(a+t,c+t)}) = 0$ and $\zeta_+(t)(C_{(a,c)}) = 0$. Then, for each $1 \le k \le K$, we have

$$\lim_{m\to\infty} \zeta_k^m(0)(B_t) = \zeta_k(0)(B_t) \qquad \text{and} \qquad \lim_{m\to\infty} \zeta_k^m(t)(B) = \zeta_k(t)(B).$$

However, by the fluid scaled version of (15) with $h = t$ and $t = 0$, for each $m \in \mathbb{M}$ and $1 \le k \le K$,

$$\begin{aligned} \zeta_k^m(t)(B) - \zeta_k^m(0)(B_t) &= \frac{1}{m} \sum_{j=Z_k^m(0)+1}^{A_k^m(t)} 1_{B_{t-t_{k,j}^m}} \left(w_{k,j}^m, p_{k,j}^m\right) \\ &= \frac{1}{m} \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A_k^m(s)}\right) \mathrm{d}E_k^m(s). \end{aligned}$$

Hence, in order to verify that (26) holds, it suffices to show that for each $1 \le k \le K$,

$$\lim_{m\to\infty} \frac{1}{m} \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A_k^m(s)}\right) \mathrm{d}E_k^m(s) = \lambda_k \int_0^t \left(\delta^+_{w(s)} \times \Gamma_k\right)(B_{t-s})\mathrm{d}s.$$

Note that for $1 \le k \le K$,

$$\begin{aligned} \lambda_k \int_0^t \left(\delta^+_{w(s)} \times \Gamma_k\right)(B_{t-s})\mathrm{d}s &= \lambda_k \int_0^t 1_{\{w(s)\ge a+t-s\}} G_k(c+t-s)\mathrm{d}s \\ &= \lambda_k \int_{\tau(a+t)\wedge t}^t G_k(c+t-s)\mathrm{d}s \\ &= \lambda_k \int_c^{c+t-\tau(a+t)\wedge t} G_k(u)\mathrm{d}u. \end{aligned}$$

Hence, in order to verify that (26) holds, it suffices to show that for each $1 \le k \le K$,

$$
\begin{aligned}
(85) \qquad & \lim_{m\to\infty} \frac{1}{m} \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&= \lambda_k \int_c^{c+t-\tau(a+t)\wedge t} G_k(u)\mathrm{d}u.
\end{aligned}
$$

In order to verify (85), fix $1 \le k \le K$. Note that for all $m \in \mathbb{M}$ and $1 \le i \le E^m_k(t)$,

$$
d_{k,i} \le p^m_{k,Z^m_k(0)+i} \le d_{k,i} + v^m_{k,i}.
$$

Then

$$
\begin{aligned}
&\int_0^t 1_{B_{t-s}} \left(w^m(s), d_{k,E^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&\qquad \le \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&\qquad \le \int_0^t 1_{B_{t-s}} \left(w^m(s), d_{k,E^m_k(s)} + v^m_{k,E^m_k(s)} \right) \mathrm{d}E^m_k(s).
\end{aligned}
$$

Let $0 < \delta < t$. Take M such that for all $m \ge M$, $\sup_{0\le s\le t} |w^m(s) - w(s)| < \delta$. Then, for all $m \ge M$,

$$
\begin{aligned}
&\int_0^t 1_{B_{t-s}} \left(w(s) - \delta, d_{k,E^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
(86) \qquad &\qquad \le \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&\qquad \le \int_0^t 1_{B_{t-s}} \left(w(s) + \delta, d_{k,E^m_k(s)} + v^m_{k,E^m_k(s)} \right) \mathrm{d}E^m_k(s).
\end{aligned}
$$

For $s \in [0,\infty)$, $w(s) - \delta \ge a + t - s$ if and only if $s \ge \tau(a+t+\delta)$. Then

$$
\begin{aligned}
&\int_0^t 1_{B_{t-s}} \left(w(s) - \delta, d_{k,E^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&\quad = \int_{\tau(a+t+\delta)\wedge t}^t 1_{[c+t-s,\infty)} \left(d_{k,E^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&\quad = \int_{\tau(a+t+\delta)\wedge t}^t 1_{[c,\infty)} \left(d_{k,E^m_k(s)} - (t-s) \right) \mathrm{d}E^m_k(s) \\
&\quad = \int_0^t 1_{[c,\infty)} \left(a^m_{k,E^m_k(s)}(t) \right) \mathrm{d}E^m_k(s)
\end{aligned}
$$

$$
\begin{aligned}
&- \int_0^{\tau(a+t+\delta)\wedge t} 1_{[c,\infty)} \left(a^m_{k,E^m_k(s)}(t) \right) \mathrm{d}E^m_k(s) \\
&\quad = \mathcal{A}^m_k(t)([c,\infty)) - \mathcal{A}^m_k(\tau(a+t+\delta)\wedge t)([c+t-\tau(a+t+\delta)\wedge t,\infty)).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\int_0^t 1_{B_{t-s}} \left(w(s)+\delta, d_{k,E^m_k(s)} + v^m_{k,E^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&\quad = \mathcal{V}^m_k(t)([c,\infty)) - \mathcal{V}^m_k(\tau(a+t-\delta)\wedge t)([c+t-\tau(a+t-\delta)\wedge t,\infty)).
\end{aligned}
$$

Then, by (47) and (86),

$$
\begin{aligned}
&\liminf_{m\to\infty} \frac{1}{m} \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A^m_k(s)} \right) \mathrm{d}E^m_k(s) \\
&\quad \geq \mathcal{R}^*_k(t)([c,\infty)) - \mathcal{R}^*_k(\tau(a+t+\delta)\wedge t)([c+t-\tau(a+t+\delta)\wedge t,\infty)).
\end{aligned}
$$

But,

$$
\begin{aligned}
&\mathcal{R}^*_k(t)([c,\infty)) - \mathcal{R}^*_k(\tau(a+t+\delta)\wedge t)([c+t-\tau(a+t+\delta)\wedge t,\infty)) \\
&\quad = \lambda_k \int_c^\infty (G_k(u) - G_k(u+t))\,\mathrm{d}u \\
&\qquad -\lambda_k \int_{c+t-\tau(a+t+\delta)\wedge t}^\infty (G_k(u) - G_k(u+\tau(a+t)\wedge t))\,\mathrm{d}u \\
&\quad = \lambda_k \int_c^{c+t-\tau(a+t+\delta)\wedge t} G_k(u)\mathrm{d}u.
\end{aligned}
$$

So then, since $\delta \in (0,t)$ is arbitrary, letting $\delta \searrow 0$ and using continuity of $\tau(\cdot)$ yields that

$$
\liminf_{m\to\infty} \frac{1}{m} \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A^m_k(s)} \right) \mathrm{d}E^m_k(s) \geq \lambda_k \int_c^{c+t-\tau(a+t)\wedge t} G_k(u)\mathrm{d}u.
$$

Similarly, using (49) in place of (47),

$$
\limsup_{m\to\infty} \frac{1}{m} \int_0^t 1_{B_{t-s}} \left(w^m(s), p^m_{k,A^m_k(s)} \right) \mathrm{d}E^m_k(s) \leq \lambda_k \int_c^{c+t-\tau(a+t)\wedge t} G_k(u)\mathrm{d}u.
$$

Hence (85) holds, as desired. □

REFERENCES

[1] Ancker, C. and Gafarian, A. (1962). Queueing with impatience customers that leave at random. *Journal of Industrial Engineering 13*, 84–90.

[2] Bacelli, F., Boyer, P., and Hebuterne, G. (1984). Single-server queue with impatient customers. *Advances in Applied Probability 16*, 887–905. MR0766784

[3] Bacelli, F. and Hebuterne, G. (1981). On queues with impatient customers. *Performance 43*, 159–179. MR0727303

[4] Billingsley, P. (1995). *Probability and Measure, 3rd Edition*. John Wiley & Sons Inc., New York. MR1324786

[5] Ethier, S. and Kurtz, T. (1986). *Markov Processes: Characterization and Convergence*. John Wiley & Sons, Inc., New York. MR0838085

[6] Folland, G. (1984). *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons Inc., New York. MR0767633

[7] Glynn, P. and Ward, A. (2005). A diffusion approximation for a GI/GI/1 queue with balking or reneging. *Queueing Systems: Theory and Applications 50*, 371–400. MR2172907

[8] Gromoll, H., Puha, A., and Williams, R. (2003). The fluid limit of a heavily loaded processor sharing queue. *Annals of Applied Probability 12*, 797–859. MR1925442

[9] Gromoll, H., Robert, P., and Zwart, B. (2008). Fluid limits for processor-sharing queues with impatience. *Mathematics of Operations Research 33*, 375–402. MR2415999

[10] Gromoll, H. and Williams, R. J. (2009). Fluid limits for networks with bandwidth sharing and general document size distributions. *The Annals of Applied Probability 19*, 243–280. MR2498678

[11] Jennings, O. and Reed, J. (2012). An overloaded multiclass FIFO queue with abandonments. *Operations Research 60*, 1282–1295. MR2998896

[12] Kallenberg, O. (1986). *Random Measures*. Academic Press, Orlando, Florida. MR0854102

[13] Kang, W. and Pang, G. (2013). Fluid limit of a many-server queueing network with abandonment. Preprint.

[14] Kang, W. and Ramanan, K. (2010). Fluid limits of many-server queues with reneging. *Annals of Applied Probability 20*, 2204–2260. MR2759733

[15] Kang, W. and Ramanan, K. (2012). Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Annals of Applied Probability 22*, 477–521. MR2953561

[16] Kaspi, H. and Ramanan, K. (2011). Law of large numbers limits for many-server queues. *Annals of Applied Probability 21*, 33–114. MR2759196

[17] Pang, G. and Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems 65*, 325–364. MR2671058

[18] Prohorov, Y. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability and its Applications 1*, 157–214. MR0084896

[19] Reed, J. and Ward, A. (2008). Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research 33*, 606–644. MR2442644

[20] Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research 54*, 37–54. MR2201245

Otis B. Jennings
323 E. Chapel Hill St
Unit 1571
Durham NC, 27702-2469
E-mail: otisj@alumni.princeton.edu

Amber L. Puha
Department of Mathematics
California State University San Marcos
333 S. Twin Oaks Valley Road
San Marcos, CA 92096-0001
E-mail: apuha@csusm.edu