MATHEMATICS OF OPERATIONS RESEARCH
Vol. 34, No. 4, November 2009, pp. 937–970
ISSN 0364-765X | EISSN 1526-5471 | 09 | 3404 | 0937

DOI 10.1287/moor.1090.0412

Law of Large Number Limits of Limited Processor-Sharing Queues

Jiheng Zhang, J. G. Dai
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332
{jrz@gatech.edu, dai@gatech.edu}

Bert Zwart
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, and Centrum voor Wiskunde en Informatica, 1098 SJ Amsterdam, The Netherlands, bertzwart@gatech.edu

Motivated by applications in computer and communication systems, we consider a processor-sharing queue where the number of jobs served is not larger than K. We propose a measure-valued fluid model for this limited processor-sharing queue and show that there exists a unique associated fluid model solution. In addition, we show that this fluid model arises as the limit of a sequence of appropriately scaled processor-sharing queues.

Key words: limited processor sharing; measure-valued process; fluid model; invariant manifold; convolution equation
MSC2000 subject classification: Primary: 60K25; secondary: 68M20, 90B22
OR/MS subject classification: Primary: queues-approximations, queues-limit theorems, queues-transient results; secondary: probability-stochastic model applications
History: Received December 4, 2007; revised August 19, 2008, April 11, 2009, and May 21, 2009. Published online in *Articles in Advance* October 7, 2009.

1. Introduction. Consider a system with a single server and an infinite capacity buffer. The server can serve up to $K \geq 1$ jobs simultaneously, distributing its attention to each of them: At any time, each job in the server is processed at a rate that is the reciprocal of the number of jobs in the server. An arriving job will immediately enter the server and start receiving service if there are less than K jobs in the server when it arrives; otherwise, it will wait in the buffer. A job will leave the system immediately after the server has fulfilled its service requirement. When the number of jobs in the server drops from K to $K-1$, the server will immediately admit the longest-waiting job from the buffer if there is any. We assume that jobs arrive according to a general arrival process, and that the job sizes are independent of each other and identically distributed.

We call this system the *limited processor-sharing queue* or LPS queue. Note that letting $K = \infty$ makes the system a *processor-sharing* (PS) queue and taking $K = 1$ reduces the system to a first in, first out FIFO queue. There is ample motivation to study this generalization. The PS discipline can be viewed as an idealization of the time-sharing protocol in computer systems, as described in Kleinrock [22] and Ritchie and Thompson [27]. The advantage is that a big job will not block the whole system as in a first-come-first-serve (FCFS) queue. However, allowing too many jobs to time share at once can lead to significant overhead (because of switching) and hence reduce overall performance. This point has already been observed in early studies of operating systems papers (Denning et al. [8], Blake [4]) as well as in more recent Web server design papers (Elnikety et al. [10], Kamra et al. [20]) and database implementation papers (Heiss and Wagner [16], Schroeder et al. [28]). So, in the modeling of many computer and communication systems, a sharing limit is normally imposed, which results in an LPS model.

There are only a few papers available that focus on performance analysis for LPS queues. An approximation for the mean response time is proposed in Avi-Itzhak and Halfin [1]. A computational analysis based on matrix geometric methods is performed in Zhang and Lipsky [31, 32]. Some stochastic ordering results are derived in Nuyens and van der Weij [24]. No rigorous analysis for general job-size distributions seems to be available. Because the model we consider is a generalization of the $G/GI/1$ PS queue and exact performance analysis of that model seems not tractable, our research focuses on obtaining limit theorems, in particular, fluid and diffusion approximations.

This paper, which characterizes the law of large number limits, constitutes the first of three major steps in our study of LPS queues. It paves the way to the study of diffusion limits in Zhang et al. [34]. The diffusion limits have nice properties and are tractable enough to obtain analytical results in the steady state. Such analytical results allow one to gain insightful approximations for various performance quantities of an LPS queue in steady state. Approximation formulas for various performance quantities are established in Zhang and Zwart [33]. In addition to providing the foundation for performance analysis, the fluid model in this paper is of independent interest.

Our study is carried out in a general setting, allowing the interarrival time and job sizes to have general distributions. Because the job-size distribution is general and multiple jobs can be in service the same time, it

is important to keep track of each of the remaining job sizes. For this purpose, we record all the remaining job sizes of all jobs in service using a measure $\mathcal{Z}(t)$ at any time t. For any Borel set $B \subset \mathbb{R}_+$, $\mathcal{Z}(t)(B)$ indicates the number of jobs in server with remaining job size belonging to B at that time. Similarly, we use a measure $\mathcal{Q}(t)$ to describe the state of the buffer, and $\mathcal{Q}(t)(B)$ indicates the number of jobs in buffer with job size belonging to B. The descriptor $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$, which takes values in the space of two dimensional vectors of Borel measures, contains a wealth of information. All the usual performance processes can be recovered from it. In fact, the measure-valued descriptor contains all the information needed to describe the dynamics of the LPS system. More details will be discussed when we give a detailed model description in §2.1. We design a set of system dynamic Equations (7) and (8) involving the descriptor for the server $\mathcal{Z}(\cdot)$ and for the buffer $\mathcal{Q}(\cdot)$. These equations are powerful enough to capture the complex dynamics and yet simple enough to perform rigorous analysis.

The framework of using measure-valued process has been successfully applied to study models where multiple jobs are processed at the same time. Existing works include Gromoll et al. [14], Puha and Williams [25], and Gromoll [12]; these papers use measure-valued descriptor for the study of PS queues that are not overloaded. Overloaded PS queues are studied in Puha et al. [26] and Jean-Marie and Robert [18]. More recently, the framework of using measure-valued processes is further developed by Gromoll and Kruk [13] and Gromoll et al. [15] in the study of PS queues with deadlines and impatience. Doytchinov et al. [9] applied a similar framework to study the earliest-deadline-first discipline. However, in most of these works, buffers are not modeled because a job immediately starts service on arrival. The only exception is Doytchinov et al. [9]; in their model, only one job is processed at a time and the buffer dynamics is described by a measure-valued process. As it will be explained in the next two paragraphs, the existence of the buffer (because of the sharing limit) creates a big challenge in our study of fluid models and the corresponding fluid limits.

To study such a complicated system, we first introduce a corresponding measure-valued fluid model. For this fluid model, we establish several fundamental properties such as existence and uniqueness of fluid model solutions. A difficulty in our study is that the fluid model involves a complicated functional equation (after some mathematical derivations including a time change) in our analysis:

$$x(u) = h(u) + \int_0^u (x(u-v) - K)^+ \, dF(v) + \rho \int_0^u (x(u-v) \wedge K) \, dF_e(v),$$

where ρ is the traffic intensity, h is a function determined by the initial condition, F is the job-size distribution and F_e is the equilibrium distribution of F (cf. see §2.1 for background and notation). In the special case of the standard PS queue, $K = \infty$ and this equation reduces to a standard renewal equation. Existence and uniqueness of the solution to a renewal equation is already known. In our case, K is finite, necessitating new methods.

We next show that the above-mentioned fluid model arises as the limit of appropriately scaled systems of LPS queues. Our analysis applies to a variety of regimes such as lightly loaded, critically loaded, and overloaded systems. When establishing convergence of a sequence of LPS queues to its fluid limit, precompactness must be proved, which turns out to be significantly more complicated than in a PS system with $K = \infty$. To put these difficulties into perspective, let $B(t)$ be the cumulative number of jobs that have entered service in $[0, t]$. The stochastic process $B = \{B(t),\ t \geq 0\}$ records the timing of jobs entering the server. For future reference, we call B the *endogenous arrival process*, the dynamics of which are much more complicated than those of the exogenous arrival process. When the system size is below K, the process B behaves like the exogenous arrival process, and when the system size is above K, the process B behaves like the departure process from the server. One of the major technical difficulties in our analysis is to show that the endogenous arrival process B is in some sense "regular." In PS queue, the process B is identical to the exogenous arrival process, whose regularity is assumed.

The regularity of the endogenous arrival process also arises in recent work of Kaspi and Ramanan [21] on multiserver queues, although our proof method is not related to theirs. Both our paper and that of Kaspi and Ramanan [21] require the assumption that the service-time distribution is continuous (although we do not need any assumption beyond continuity, as in Kaspi and Ramanan [21]). Note that the LPS queue is similar to the FCFS system with K servers. The difference is that the service rate in many a server queue is equal to one while it fluctuates in an LPS queue. On the other hand, the many-server queue is not workload conserving but the LPS queue is.

This paper is organized as follows. In §2, we give a detailed description of the LPS queue and its fluid analogue. The main results of this paper are presented in §3. Section 4 investigates properties of the fluid model. Convergence of a scaled sequence of systems is considered in §§5 and 6. Precompactness is established in §5 and §6 shows that every limit point is a fluid model solution.

1.1. Notation. The following notation will be used throughout. Let $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{R}$ denote the set of natural numbers, integers, and real numbers, respectively. Let $\mathbb{R}_+ = [0, \infty)$. For $a, b \in \mathbb{R}$, write a^+ for the positive part of a, $\lfloor a \rfloor$ for the integer part, $\lceil a \rceil$ for $\lfloor a \rfloor + 1$, $a \vee b$ for the maximum, and $a \wedge b$ for the minimum.

Let $\mathbf{M}_1$ and $\mathbf{M}_2$ denote the set of all nonnegative finite Borel measures on $[0, \infty)$ and $(0, \infty)$, respectively. To simplify the notation, let us take the convention that for any Borel set $A \subset \mathbb{R}$, $\nu(A \cap (-\infty, 0)) = 0$ for any $\nu \in \mathbf{M}_1$ and $\nu(A \cap (-\infty, 0]) = 0$ for any $\nu \in \mathbf{M}_2$. Also, by this convention, $\mathbf{M}_2$ is embedded as a subspace of $\mathbf{M}_1$. For $\nu_1, \nu_2 \in \mathbf{M}_1$, the Prohorov metric is defined to be

$$\mathbf{d}[\nu_1, \nu_2] = \inf\{\epsilon > 0\colon \nu_1(A) \le \nu_2(A^\epsilon) + \epsilon \text{ and } \nu_2(A) \le \nu_1(A^\epsilon) + \epsilon \text{ for all closed Borel set } A \subset \mathbb{R}_+\},$$

where $A^\epsilon = \{b \in \mathbb{R}_+\colon \inf_{a \in A} |a - b| < \epsilon\}$. This is the metric that induces the topology of weak convergence of finite Borel measures. (See §6 in Billingsley [3].) For any Borel measurable function $g\colon \mathbb{R}_+ \to \mathbb{R}$, the integration of this function with respect to the measure $\nu \in \mathbf{M}_1$ is denoted by $\langle g, \nu \rangle$.

Let $\mathbf{M}_1 \times \mathbf{M}_2$ denote the Cartesian product. There are a number of ways to define the metric on the product space. For convenience, we define the metric to be the maximum of the Prohorov metric between each component. With a little abuse of notation, we still use $\mathbf{d}$ to denote this metric.

Let $(\mathbf{E}, \pi)$ be a general metric space. We consider the space $\mathbf{D}$ of all right-continuous $\mathbf{E}$-valued functions with finite left limits defined either on a finite interval $[0, T]$ or the infinite interval $[0, \infty)$. We refer to the space as $\mathbf{D}([0, T], \mathbf{E})$ or $\mathbf{D}([0, \infty), \mathbf{E})$, depending on the function domain. The space $\mathbf{D}$ is also known as the space of *càdlàg* functions. For $g(\cdot), g'(\cdot) \in \mathbf{D}([0, T], \mathbf{E})$, the uniform metric is defined as

$$\upsilon_T[g, g'] = \sup_{0 \le t \le T} \boldsymbol{\pi}[g(t), g'(t)]. \tag{1}$$

However, a more useful metric that we will use is the following Skorohod J_1 metric:

$$\varrho_T[g, g'] = \inf_{f \in \Lambda_T} (\|f\|_T^\circ \vee \upsilon_T[g, g' \circ f]), \tag{2}$$

where $g \circ f(t) = g(f(t))$ for $t \ge 0$ and Λ_T is the set of strictly increasing and continuous mapping of $[0, T]$ onto itself and

$$\|f\|_T^\circ = \sup_{0 \le s < t \le T} \left| \log \frac{f(t) - f(s)}{t - s} \right|.$$

If $g(\cdot)$ and $g'(\cdot)$ are in the space $\mathbf{D}([0, \infty), \mathbf{E})$, the Skorohod J_1 metric is defined as

$$\varrho[g, g'] = \int_0^\infty e^{-T} (\varrho_T[g, g'] \wedge 1)\, dT. \tag{3}$$

By saying convergence in the space $\mathbf{D}$, we mean the convergence under the Skorohod J_1 topology, which is the topology induced by the Skorohod J_1 metric (Ethier and Kurtz [11]).

We use "$\to$" to denote the convergence in a general metric space $(\mathbf{E}, \boldsymbol{\pi})$ and use "$\Rightarrow$" to denote the convergence in distribution of random variables taking value in the metric space $(\mathbf{E}, \boldsymbol{\pi})$.

2. The LPS queue and dynamic equations.

In this section, we first give a detailed description of the stochastic process associated with the LPS queue, and then define a corresponding fluid model that serves as an important tool to study the stochastic process.

2.1. Stochastic model. We consider a $G/GI/1$ queue operated under the limited processor-sharing policy with the sharing limit equal to K. We use $Q(t)$, $Z(t)$, and $X(t)$ to denote the number of jobs in the buffer, number of jobs in service, and the total number of jobs in the system at time t, respectively. Thus,

$$X(t) = Q(t) + Z(t) \quad \text{for } t \ge 0.$$

The system is allowed to be nonempty initially, i.e., $X(0) > 0$. We index jobs by $i = -X(0) + 1, -X(0) + 2, \ldots, 0, 1, \ldots$. The first $X(0)$ jobs are initially in the system, with jobs $i = -X(0) + 1, \ldots, -Q(0)$ in service and jobs $i = -Q(0) + 1, \ldots, 0$ waiting in the buffer. Jobs arrived after time 0 are indexed by $i = 1, 2 \ldots$. Let $E(t)$ denote the number of jobs that arrive to the system during time interval $(0, t]$ for all $t \ge 0$. According to the policy, a job may have to wait for a certain amount of time after arrival to get service. Let w_i denote the

waiting time and U_i denote the arrival time of the ith job for all $i > -X(0)$. By convention, $U_i = 0$ for $i < 0$ and $w_i = 0$ for $i \le -Q(0)$. Let

$$\tau_i = U_i + w_i, \quad i > -X(0).$$

The quantity τ_i can be viewed as the time that the ith job starts service. We use v_i to denote the job size of the ith job for all $i > -Q(0)$. We assume that $\{v_i\}_{i=-\infty}^{\infty}$ is a sequence of i.i.d. random variables with distribution F. For jobs with index $-X(0) < i \le -Q(0)$, i.e., the first $Z(0)$ jobs that are initially in service, we use $\tilde{v}_i$ to denote the job sizes of these jobs. We call $\{E(\cdot), \{v_i\}_{i=1}^{\infty}\}$ the stochastic primitives of the system and $\{Z(0), Q(0), \{v_i\}_{i=-\infty}^{0}, \{\tilde{v}_i\}_{i=-\infty}^{0}\}$ the initial conditions of the system.

Next, we introduce a measure-valued state descriptor $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$, which describes the evolution of the system with given initial conditions and stochastic primitives. Let $\mathcal{Q}(\cdot)$ and $\mathcal{Z}(\cdot)$ be $\mathbf{M}_1$-valued and $\mathbf{M}_2$-valued stochastic processes, respectively. For any Borel set $A \subset [0, \infty)$, $\mathcal{Q}(t)(A)$ denotes the total number of jobs in buffer whose job size belongs to A, and for any Borel set $A \subset (0, \infty)$, $\mathcal{Z}(t)(A)$ denotes the total number of jobs in service whose residual job size belongs to set A. Note that here we distinguish the spaces for buffer and server descriptors. The reason is that we allow jobs with size zero to arrive and wait in the buffer. However, a job in service will immediately leave the system once its remaining service time becomes zero. So no job in service can have zero remaining service time. It is clear that we have the following relationship:

$$Q(t) = \langle 1, \mathcal{Q}(t) \rangle, \qquad Z(t) = \langle 1, \mathcal{Z}(t) \rangle.$$

Define the *cumulative service amount* up to time t by

$$S(t) = \int_0^t \psi(Z(\tau))\, d\tau, \tag{4}$$

where $\psi(x) = 1/x$ if $x > 0$ and $\psi(x) = 0$ if $x = 0$. A job will have received a cumulative amount of processing time

$$S(s, t) = \int_s^t \psi(Z(\tau))\, d\tau$$

during time interval $[s, t]$ if it is in service in this time period. Let

$$B(t) = E(t) - Q(t). \tag{5}$$

Note that at time $t \ge 0$, $B(t)$ is the index of the last job that has entered into service by time t. Thus

$$B(s, t) = B(t) - B(s) \tag{6}$$

represents the number of jobs that have entered the server during time interval $(s, t]$. Using the notation introduced in this section, the state descriptor can be written as

$$\mathcal{Q}(t)(A') = \sum_{i=B(t)+1}^{E(t)} \delta_{v_i}(A'), \tag{7}$$

$$\mathcal{Z}(t)(A) = \sum_{i=-X(0)+1}^{-Q(0)} \delta_{\tilde{v}_i}(A + S(t)) + \sum_{i=-Q(0)+1}^{B(t)} \delta_{v_i}(A + S(\tau_i, t)) \tag{8}$$

for any Borel sets $A' \subseteq [0, \infty)$ and $A \subseteq (0, \infty)$ and $t \ge 0$, where δ_a denotes the Dirac measure of point a on $\mathbb{R}$ and $A + y = \{a + y\colon a \in A\}$. Because of the LPS policy, the sharing limit K must be enforced at any time t,

$$Q(t) = (X(t) - K)^+, \tag{9}$$

$$Z(t) = (X(t) \wedge K). \tag{10}$$

We call (7) and (8) the *stochastic dynamic equations* and (9) and (10) the policy constraints.

Another performance process that we are interested in is $D(t)$, the number of jobs that have left the system in $(0, t]$. It can be written as

$$D(t) = X(0) + E(t) - X(t). \tag{11}$$

For $t \ge 0$, the workload of the system $W(t)$ is defined to be the amount of time that the server remains busy *if* no more arrivals are allowed into the system at time t. Using state descriptor $(\mathcal{Q}, \mathcal{Z})$, we can recover the workload $W(t)$ at time $t > 0$ by

$$W(t) = \langle \chi, \mathcal{Q}(t) + \mathcal{Z}(t) \rangle, \tag{12}$$

where χ denotes the identity function on $\mathbb{R}$.

2.2. Fluid model. In this section, we propose a fluid analogue of the LPS system. Given a measure-valued process $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M}_1 \times \mathbf{M}_2)$, for $t \geq 0$, let

$$\bar{Q}(t) = \langle 1, \bar{\mathcal{Q}}(t)\rangle, \tag{13}$$

$$\bar{Z}(t) = \langle 1, \bar{\mathcal{Z}}(t)\rangle, \tag{14}$$

$$\bar{X}(t) = \bar{Q}(t) + \bar{Z}(t), \tag{15}$$

$$\bar{B}(t) = \lambda t - \bar{Q}(t), \tag{16}$$

$$\bar{D}(t) = \lambda t + \bar{X}(0) - \bar{X}(t), \tag{17}$$

where λ is a positive constant that is interpreted as the arrival rate. These quantities are the fluid analogues of $Q(t), Z(t), B(t), D(t)$, and $X(t)$ in the stochastic model. Let ν be the probability measure associated with the job-size distribution F. We call ν the job-size measure. Denote $\beta = \langle \chi, \nu\rangle$ the mean of the job size, and define

$$\rho = \lambda\beta$$

to be the *traffic intensity* of the LPS queue. Define the *fluid cumulative service amount* up to time t by

$$\bar{S}(t) = \int_0^t \phi_\rho(\bar{Z}(\tau))\, d\tau, \tag{18}$$

where $\phi_\rho(x) = 1/x$ for all x, $\rho > 0$ and

$$\phi_\rho(0) = \begin{cases} \infty & \rho \in (0,1], \\ 0 & \rho \in (1,\infty). \end{cases} \tag{19}$$

For $0 \leq s \leq t$, denote

$$\bar{S}(s,t) = \int_s^t \phi_\rho(\bar{Z}(\tau))\, d\tau. \tag{20}$$

This is how the fluid cumulative service amount is defined, and it turns out that this definition serves the purpose of studying the fluid model very well. Here, we give some intuitive explanation of why using the function ϕ_ρ instead of ψ in (4). In the corresponding stochastic process, when there is no job in the system, the server idles, implying $\psi(0) = 0$. In the fluid model with $\rho \leq 1$, intuitively, the amount of fluid in service $\bar{Z}(\cdot)$ will stay at zero once it reaches zero. Because fluids flow in at a constant rate λ, the server, instead of idling, actually finishes service immediately when an infinitesimal amount of fluid enters service. Thus, very naturally, $\phi_\rho(0) = \infty$ when $\rho \leq 1$. However, when $\rho > 1$, intuitively, the queue size should grow if starts at zero. To rule out the solution $z(\cdot) \equiv 0$, we define $\phi_\rho(0) = 0$. Note that the definitions of fluid model solutions for the standard PS queue also depend on the load (cf. Gromoll et al. [14], Puha et al. [26]).

An element $(\xi, \mu) \in \mathbf{M}_1 \times \mathbf{M}_2$ is called a *valid initial condition* if

$$\xi = (\langle 1, \xi + \mu\rangle - K)^+ \nu,$$

$$\langle 1, \mu\rangle = \langle 1, \xi + \mu\rangle \wedge K.$$

Roughly speaking, validity of an initial state means that the initial state is consistent with the limited sharing policy and initial waiting jobs have the same service distribution as arriving jobs. Denote $\mathcal{I}$ the set of all valid initial conditions.

We now introduce the following *fluid dynamic equations*, which are analogous to (7) and (8). For all $t \geq 0$ and $A_y = (y,\infty)$ with $y \geq 0$,

$$\bar{\mathcal{Q}}(t)(A_y) = \xi(A_y) + (\bar{Q}(t) - \bar{Q}(0))\nu(A_y), \tag{21}$$

$$\bar{\mathcal{Z}}(t)(A_y) = \mu(A_y + \bar{S}(t)) + \int_0^t \nu(A_y + \bar{S}(s,t))\, d\bar{B}(s), \tag{22}$$

where $\bar{Q}(\cdot)$, $\bar{Z}(\cdot)$, $\bar{X}(\cdot)$, $\bar{B}(\cdot)$, and $\bar{S}(\cdot)$ are defined in (13)–(20). They are subject to the following constraints:

$$\bar{B}(\cdot) \text{ is nondecreasing}, \tag{23}$$

$$\bar{Q}(\cdot) = (\bar{X}(\cdot) - K)^+, \tag{24}$$

$$\bar{Z}(\cdot) = (\bar{X}(\cdot) \wedge K). \tag{25}$$

The above equations define a fluid model, which we denote by the triple (K, λ, ν).

DEFINITION 2.1. $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M}_1 \times \mathbf{M}_2)$ is a solution to the fluid model (K, λ, ν) with a valid initial state (ξ, μ) if it satisfies the fluid dynamic Equations (21) and (22), subject to the constraints (23)–(25).

Similar to the stochastic model, the fluid workload $\bar{W}(t)$ at any time $t > 0$ is defined as

$$\bar{W}(t) = \langle \chi, \bar{\mathcal{Q}}(t) + \bar{\mathcal{Z}}(t) \rangle. \tag{26}$$

3. Main results. This section presents the main results of this paper. Our first set of results is concerned with several key properties of fluid model solutions such as existence, uniqueness, and stability. Our second set of results shows that the fluid model arises as the limit of an appropriately scaled sequence of stochastic LPS systems.

3.1. Properties of fluid model solutions. We first state several key properties of our fluid model solution. The following theorem establishes the existence and uniqueness of the fluid model solution.

THEOREM 3.1. *Assume that the job-size measure ν satisfies*

$$\langle \chi, \nu \rangle < \infty, \tag{27}$$

$$\nu(\{0\}) = 0. \tag{28}$$

There exists a unique solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ to the fluid model (K, λ, ν) with initial condition $(\xi, \mu) \in \mathcal{I}$.

We have the following workload-conserving property for any fluid model solution.

PROPOSITION 3.1. *Assume that the job-size measure ν satisfies (27) and (28). The fluid workload $\bar{W}(\cdot)$ of any solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ to the fluid model (K, λ, ν) with initial condition $(\xi, \mu) \in \mathcal{I}$ satisfies*

$$\bar{W}(t) = (\langle \chi, \xi + \mu \rangle + (\rho - 1)t)^+ \quad \text{for all } t \geq 0.$$

We now turn to stability properties of our fluid model. Although the results are intuitively clear, the stability properties of fluid model solutions require formal proof in the measure-valued setup. The following definitions are analogous to the standard fluid model as in Dai [6, 7].

DEFINITION 3.1. A fluid model (λ, K, ν) is *weakly stable* if any fluid model solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ with initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$ satisfies $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$ for all $t \geq 0$.

A fluid model (λ, K, ν) is *stable* if for any initial condition $(\xi, \mu) \in \mathcal{I}$ satisfying $0 < w = \langle \chi, \xi + \mu \rangle < \infty$, there exists a finite time δ (only depending on w) such that any fluid model solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ with this initial condition satisfies $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$ for all $t \geq \delta$.

THEOREM 3.2. *Assume that the job-size measure ν satisfies (27) and (28). A fluid model (λ, K, ν) is weakly stable if the traffic intensity $\rho \leq 1$; it is stable if the traffic intensity $\rho < 1$.*

We prove Theorem 3.1, Proposition 3.1, and Theorem 3.2 in §4.

3.2. Fluid model as fluid limit. The main motivation to study the fluid model is that it serves as the weak law of large number limits of the stochastic process described in §2.1. Consider a sequence of limited processor-sharing queues indexed by r, where r increases to ∞ through a sequence in $(0, \infty)$. Each model is defined in the same way as in §2.1. To distinguish models with different indices, quantities of the rth model are accompanied by superscript r. Each model may be defined on a different probability space $(\Omega^r, \mathcal{F}^r, \mathbb{P}^r)$. Our results concern the asymptotic behavior of the descriptor under the *fluid* scaling, which is defined by

$$\bar{\mathcal{Q}}^r(t) = \frac{1}{r}\mathcal{Q}^r(rt), \qquad \bar{\mathcal{Z}}^r(t) = \frac{1}{r}\mathcal{Z}^r(rt) \tag{29}$$

for all $t \geq 0$. We are also interested in fluid-scaled versions of other quantities like the workload and queue length processes. Note that $\bar{Q}^r(\cdot)$, $\bar{Z}^r(\cdot)$, and $\bar{W}^r(\cdot)$ are actually functions of $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$, so the scaling for these quantities is defined as the functions of the corresponding scaling for $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$, i.e.,

$$\bar{Q}^r(t) = \langle 1, \bar{\mathcal{Q}}^r(t) \rangle = \frac{1}{r}Q^r(rt), \tag{30}$$

$$\bar{Z}^r(t) = \langle 1, \bar{\mathcal{Z}}^r(t) \rangle = \frac{1}{r}Z^r(rt), \tag{31}$$

$$\bar{W}^r(t) = \langle \chi, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t) \rangle = \frac{1}{r}W^r(rt) \tag{32}$$

for all $t \geq 0$. Similarly, we define the fluid scaling for cumulative service amount $S^r(s,t)$ to be

$$\bar{S}^r(s,t) = \int_s^t \psi(\bar{Z}^r(\tau))\, d\tau \tag{33}$$

for $0 \leq s \leq t$. The fluid scaling for the external arrival process is defined as

$$\bar{E}^r(t) = \frac{1}{r} E^r(rt). \tag{34}$$

It follows from (5) and (11) that the scaling for $\bar{B}^r(\cdot)$ and $\bar{D}^r(\cdot)$ should be defined by

$$\bar{B}^r(t) = \frac{1}{r} B^r(rt), \qquad \bar{D}^r(t) = \frac{1}{r} D^r(rt) \tag{35}$$

for all $t \geq 0$.

To establish results on convergence of the above sequence of stochastic processes, we need the following conditions that are quite general and standard. We assume that the arrival processes satisfy

$$\bar{E}^r(\cdot) \;\Rightarrow\; \lambda \cdot \quad \text{as } r \to \infty, \tag{36}$$

where λ is a positive constant. The job-size measures ν^r satisfy that as $r \to \infty$

$$\mathbf{d}[\nu^r, \nu] \to 0, \tag{37}$$

$$\langle \chi^{1+p}, \nu^r \rangle \to \langle \chi^{1+p}, \nu \rangle < \infty \quad \text{for some } p > 0, \tag{38}$$

where ν satisfies

$$\nu \text{ has no atoms.} \tag{39}$$

The law of large number scaling speeds up the processes r times, so we need to scale the sharing limit accordingly:

$$\lim_{r\to\infty} K^r/r \to K > 0. \tag{40}$$

Also, the following initial condition will be assumed:

$$(\bar{\mathcal{Q}}^r(0), \bar{\mathcal{Z}}^r(0)) \;\Rightarrow\; (\xi^*, \mu^*), \tag{41}$$

$$\langle \chi^{1+p}, \bar{\mathcal{Q}}^r(0) + \bar{\mathcal{Z}}^r(0) \rangle \;\Rightarrow\; \langle \chi^{1+p}, \xi^* + \mu^* \rangle, \tag{42}$$

where p is the same as in (38) and (ξ^*, μ^*) is a deterministic element in $\mathcal{I}$ and

$$\mu^* \text{ has no atoms.} \tag{43}$$

The following proposition is a well-known result for a single server queue operating under a nonidling service discipline. Readers are referred to §5 in Gromoll et al. [14] for a proof.

PROPOSITION 3.2. *Assume the sequence of LPS queues satisfies* (36)–(42). *As* $r \to \infty$, *we have*

$$\bar{W}^r(\cdot) \;\Rightarrow\; \bar{W}(\cdot),$$

where $\bar{W}(t) = (\langle \chi, \xi^* + \mu^* \rangle + (1-\rho)t)^+$ *for all* $t \geq 0$.

Because the LPS is also a nonidling service discipline, the above limit of the workload process still holds for our model.

However, the limiting of the job-size process and many other performance processes as introduced above is far from clear. Our main result establishes the fluid limit of the measure-valued processes (Theorem 3.3) from which the fluid limit of many interesting performance processes follows directly (Corollary 3.1).

THEOREM 3.3. *If the sequence of limited processor-sharing queues satisfies* (36)–(43), *then*

$$(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot)) \;\Rightarrow\; (\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \quad \text{as } r \to \infty,$$

where $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ *is the unique solution to the fluid model* (K, λ, ν) *with initial condition* (ξ^*, μ^*).

Because all performance measures can be recovered from the descriptor $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$ through continuous mappings, we have the following corollary.

COROLLARY 3.1. *Assume the sequence of limited processor queues satisfies* (36)–(43). *As* $r \to \infty$, *we have*

$$(\bar{Q}^r(\cdot), \bar{Z}^r(\cdot), \bar{B}^r(\cdot), \bar{D}^r(\cdot)) \;\Rightarrow\; (\bar{Q}(\cdot), \bar{Z}(\cdot), \bar{B}(\cdot), \bar{D}(\cdot)),$$

where $\bar{Q}(\cdot)$, $\bar{Z}(\cdot)$, $\bar{B}(\cdot)$, $\bar{D}(\cdot)$ *are as defined in* (13)–(17).

Corollary 3.1 follows immediately from Theorem 3.3. We omit the proof for brevity. We will prove Theorem 3.3 in §6.

4. Properties of fluid model solutions. Note that the fluid amount of jobs in service $\bar{Z}(t) = \bar{\mathfrak{Z}}(t)(A_0)$ for all $t \geq 0$. (Recall that $A_y = (y, \infty)$ for all $y \geq 0$.) By (22) in Definition 2.1, we have

$$\bar{Z}(t) = \mu(A_{\bar{S}(t)}) + \int_0^t [1 - F(\bar{S}(s,t))]\, d\bar{B}(s). \tag{44}$$

To further analyze the fluid model, we need to distinguish between different cases. We first consider the case where the initial condition is nonzero. In this case, there exists a nontrivial interval on which the server size never reaches zero. Thus, we can do a time change to obtain the Equation (50), which is the key equation in our analysis. Through this analysis, we can characterize the fluid model solution on a small interval. We then use the "restarting" lemma (Lemma 4.2) to extend the result to a larger interval. After that case, we consider the case where the initial condition is zero and traffic intensity $\rho \leq 1$. Basically, we show that the fluid model solution will stay at zero. Finally, we study the case with zero initial condition and $\rho > 1$. Briefly speaking, the fluid model solution will grow "linearly" in this case.

4.1. Starting with a nonzero valid initial condition. If the valid initial condition $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$, then $Z(0) = \langle 1, \mu \rangle > 0$. Let

$$t^* = \inf\{s > 0\colon \bar{Z}(s) = 0\}. \tag{45}$$

Because $\bar{Z}(0) > 0$, by right continuity of $\bar{Z}(\cdot)$ we have $t^* > 0$. The following algebra will be performed on the interval $[0, t^*)$, where the function $\bar{S}(\cdot)$ as defined in (18) has an inverse, which is denoted by $\bar{T}(\cdot)$. By the implicit function theorem,

$$\bar{T}'(v) = \bar{Z}(\bar{T}(v)). \tag{46}$$

Performing the change of variables $u = \bar{S}(t)$ and $v = \bar{S}(s)$ to (44), we get

$$\bar{Z}(\bar{T}(u)) = \mu(A_u) + \lambda \int_0^u [1 - F(u-v)]\bar{Z}(\bar{T}(v))\, dv - \int_0^u [1 - F(u-v)]\, d\bar{Q}(\bar{T}(v)).$$

Through the change of variable $v \leftarrow u - v$ and integration by parts, we obtain

$$\begin{aligned}\bar{Z}(\bar{T}(u)) &= \mu(A_u) + \lambda\beta \int_0^u \bar{Z}(\bar{T}(u-v))\, dF_e(v) - [1 - F(0)]\bar{Q}(\bar{T}(u)) \\ &\quad + [1 - F(u)]\bar{Q}(0) + \int_0^u \bar{Q}(\bar{T}(u-v))\, dF(v),\end{aligned}$$

where F_e is the equilibrium distribution of F that can be written as $F_e(x) = (1/\beta)\int_0^x [1 - F(y)]\, dy$. By condition (28), $F(0) = 0$. Now, we obtain the key relationship

$$\bar{Q}(\bar{T}(u)) + \bar{Z}(\bar{T}(u)) = \xi(A_u) + \mu(A_u) + \int_0^u \bar{Q}(\bar{T}(u-v))\, dF(v) + \rho \int_0^u \bar{Z}(\bar{T}(u-v))\, dF_e(v) \tag{47}$$

for all $0 \leq u < u^* = \bar{S}(t^*)$. To simplify notation, denote

$$h(u) = \xi(A_u) + \mu(A_u), \tag{48}$$

$$x(u) = q(u) + z(u), \tag{49}$$

where $q(u) = \bar{Q}(\bar{T}(u))$ and $z(u) = \bar{Z}(\bar{T}(u))$. By (24) and (25), the above equation can be written as

$$x(u) = h(u) + \int_0^u (x(u-v) - K)^+\, dF(v) + \rho \int_0^u (x(u-v) \wedge K)\, dF_e(v). \tag{50}$$

This functional equation would simplify to a renewal equation if $K = \infty$ or $K = 0$. (It should be pointed out that the special cases $K = \infty$ and $K = 0$ correspond to PS queue and FIFO queue, respectively. The fluid model is proved to be the limit of a sequence of fluid-scaled processes under several conditions including (40), i.e., $K = \lim_{r\to\infty} K^r/r$. In the PS queue, each $K^r = \infty$, so $K = \infty$; in the FIFO queue, each $K^r = 1$, so $K = 0$. The latter represents the fact that for FIFO queue, the profile of server is washed away in fluid scaling. In fact, although the fluid model in earlier works on the PS queue (Gromoll et al. [14], Gromoll [12]) or related models (Gromoll and Kruk [13], Gromoll et al. [15]) is defined in a different way, the mathematical analysis is essentially focused on Equation (50) with $K = \infty$, which is a renewal equation. However, when $0 < K < \infty$, Equation (50) is no longer a renewal equation.

We provide a general tool to study the integral Equation (50) in Appendix A. The tool represents one of our major technical contributions of this paper. Lemma A.1 in Appendix A requires even weaker conditions than we need, which will be useful for future purposes. In our setting, condition (28) and the definition of $h(\cdot)$ in (48) imply that all the conditions needed in Lemma A.1 are satisfied. Building on Lemma A.1, we establish the existence and uniqueness of fluid model solutions on a small interval through Lemma 4.1.

LEMMA 4.1. *Assume (27) and (28). For any nonzero initial condition $(\xi, \mu) \in \mathcal{I}$, there exists a $t' > 0$ such that the fluid model (K, λ, ν) has a unique solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ on $[0, t']$ satisfying the initial condition and*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) \neq (\mathbf{0}, \mathbf{0}) \quad \text{for all } t \in [0, t').$$

PROOF. Lemma A.1 establishes the uniqueness and existence of solution to (50) on the interval $[0, a]$, where a is positive and does not depend on initial condition. Let

$$a' = \inf\{u \le a\colon x(u) = 0\}. \tag{51}$$

We have that $a' > 0$ because $x(\cdot)$ is right continuous and the initial condition is nonzero. Now, let

$$\bar{T}(u) = \int_0^u (x(v) \wedge K)\, dv.$$

It is clear that $\bar{T}(\cdot)$ is differentiable and strictly increasing on $[0, a']$. Let $\bar{S}(t)$ denote its inverse function, which is still differentiable and strictly increasing on $[0, a']$. Now, define

$$\bar{X}(t) = x(\bar{S}(t))$$

and $\bar{Q}(t) = (\bar{X}(t) - K)^+$, $\bar{Z}(t) = \bar{X}(t) \wedge K$. Because $x(\cdot)$ is *càdlàg* and $\bar{T}(\cdot)$ is continuous, it is clear that $\bar{X}(\cdot)$ is *càdlàg*. By the implicit function theorem,

$$\bar{S}'(t) = \frac{1}{T'(\bar{S}(t))} = \frac{1}{\bar{X}(t) \wedge K} = \frac{1}{\bar{Z}(t)}.$$

Because $x(\cdot)$ is a solution to (50) on the interval $[0, a']$, $\bar{X}(\cdot)$ is a solution to (47) (and thus to (44)) on the interval $[0, t']$, where

$$t' = \bar{T}(a'). \tag{52}$$

Let $\bar{B}(t) = \lambda t - \bar{Q}(t)$ for all $t \in [0, t']$. Because (ξ, μ) is a valid initial condition,

$$\xi([0, u]) = (\langle 1, \xi + \mu\rangle - K)^+ F(u) \qquad \text{and} \qquad \langle 1, \mu\rangle = \langle 1, \xi + \mu\rangle \wedge K.$$

Because $\mu \neq \mathbf{0}$, let $G(\cdot) = \mu([0, \cdot])/\langle 1, \mu\rangle$, which is a distribution function. By the definition of $h(\cdot)$ in (48), we have that $h(u) = (h(0) \wedge K)[1 - G(u)] + (h(0) - K)^+[1 - F(u)]$. Thus, it satisfies the conditions in Lemma A.2. So, by Lemma A.2, $\bar{B}(\bar{T}(u))$ is nondecreasing on the interval $[0, a']$. Thus, $\bar{B}(t)$ is nondecreasing on the interval $[0, t']$ because $\bar{T}(u)$ is strictly increasing on $[0, a']$. Define

$$\bar{\mathcal{Q}}(t)(A_y) = \bar{Q}(t)[1 - F(y)],$$

$$\bar{\mathcal{Z}}(t)(A_y) = \mu(A_y + \bar{S}(t)) + \int_0^t \nu(A_y + \bar{S}(s, t))\, d\bar{B}(s),$$

where $\bar{S}(s, t)$ is defined in (20). This only defines $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ for Borel sets of the form (y, ∞). By Dynkin's π-λ theorem (cf. Theorem 3.2 in Billingsley [2]), it defines the measure for all Borel sets in $(0, \infty)$. It is clear by the first equation that $\bar{Q}(t) = \langle 1, \bar{\mathcal{Q}}(t)\rangle$. Plug A_0 in both sides of the second equation above to get

$$\begin{aligned}\langle 1, \bar{\mathcal{Z}}(t)\rangle &= \bar{\mathcal{Z}}(t)(A_0)\\ &= \mu(A_{\bar{S}(t)}) + \int_0^t [1 - F(\bar{S}(s, t))]\, d\bar{B}(s)\\ &= \bar{Z}(t),\end{aligned}$$

where the last equality is because of (44). So $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies the definition of a fluid model solution, implying the existence. The measure $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ will never be zero on $[0, t')$ because of (51) and (52).

To prove uniqueness, assume there is another solution $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$ for the same initial condition. By Definition 2.1, it must satisfy (21)–(25). Let

$$t^\dagger = \inf\{t \geq 0\colon \bar{X}^\dagger(t) > 0\}.$$

We know that $t^\dagger > 0$ by right continuity of $\bar{X}^\dagger(t)$ and the nonzero initial condition. Thus, $\bar{S}^\dagger(\cdot)$ has inverse $\bar{T}^\dagger(\cdot)$ on $[0, t^\dagger]$. Let

$$x^\dagger(u) = \bar{X}^\dagger(\bar{T}^\dagger(u)) \quad \text{for } 0 \leq u \leq \bar{S}^\dagger(t^\dagger).$$

By (21)–(25), $x^\dagger(\cdot)$ must satisfy (50) on $[0, \bar{S}^\dagger(t^\dagger)]$. Because of the the uniqueness of solutions to (50),

$$x^\dagger(u) = x(u) \quad \text{for } u \leq \min(\bar{S}^\dagger(t^\dagger), a').$$

We first claim that $\bar{S}^\dagger(t^\dagger) \geq a'$. Otherwise, $\bar{S}^\dagger(t^\dagger) < a' \leq a$. By (51),

$$\bar{X}^\dagger(t^\dagger) = x^\dagger(\bar{S}^\dagger(t^\dagger)) = x(\bar{S}^\dagger(t^\dagger)) > 0,$$

which contradicts the definition of $t^\dagger$. Thus, $x^\dagger(\cdot)$ and $x(\cdot)$ agree on the interval $[0, a']$, which implies that

$$\frac{d}{du}\bar{T}(u) = x(u) \wedge K = x^\dagger(u) \wedge K = \frac{d}{du}\bar{T}^\dagger(u).$$

Because both $\bar{T}(u)$ and $\bar{T}^\dagger(u)$ are absolutely continuous, $\bar{T}^\dagger(u) = \bar{T}(u)$ for all $u \leq a'$. This means that $\bar{X}^\dagger(t) = \bar{X}(t)$ and $\bar{S}^\dagger(t) = \bar{S}(t)$ for all $t \leq t'$. By (21) and (22), $(\bar{\mathcal{Q}}^\dagger(t), \bar{\mathcal{Z}}^\dagger(t)) = (\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t))$ for all $t \leq t'$. Uniqueness is proved. □

Thus far, we have established the existence and uniqueness of fluid model solution on a nontrivial interval $[0, t']$. The following "restarting" lemma helps to extend the result in Lemma 4.1 to a larger interval.

LEMMA 4.2. *Assume (27) and (28). Let $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot))$ be a solution to the fluid model (K, λ, ν) on the interval $[0, t_1]$ for some $t_1 > 0$. If $(\bar{\mathcal{Q}}_2(\cdot), \bar{\mathcal{Z}}_2(\cdot))$ is a solution to the fluid model with initial condition $(\bar{\mathcal{Q}}_1(t_1), \bar{\mathcal{Z}}_1(t_1))$ on the interval $[0, t_2]$ for some $t_2 > 0$, then $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is a fluid model solution on $[0, t_1 + t_2]$, where*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = \begin{cases} (\bar{\mathcal{Q}}_1(t), \bar{\mathcal{Z}}_1(t)) & \text{if } t \in [0, t_1], \\ (\bar{\mathcal{Q}}_2(t_1 + t), \bar{\mathcal{Z}}_2(t_1 + t)) & \text{if } t \in [t_1, t_1 + t_2]. \end{cases}$$

PROOF. The proof of Lemma 4.2 is very straightforward. It is clear that $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies the fluid dynamic equations on the interval $[0, t_1]$. For any $t \in (t_1, t_1 + t_2]$, plugging t_1 and $t_1 + (t - t_1)$ into (21) and (22) and then taking the summation gives

$$\bar{\mathcal{Q}}(t)(A_y) = \bar{\mathcal{Q}}(0)(A_y) + [\bar{Q}(t) - \bar{Q}(0)]\nu(A_y),$$

$$\bar{\mathcal{Z}}(t)(A_y) = \bar{\mathcal{Z}}(0)(A_y + \bar{S}(t) - \bar{S}(0)) + \int_0^t \nu(A_y + \bar{S}(t) - \bar{S}(s))\, d[\lambda s - \bar{Q}(s)]$$

for all $A_y = (y, \infty)$, $y \geq 0$. Thus, $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ satisfies the fluid dynamic equations on the interval $[0, t_1 + t_2]$. Clearly, it also satisfies all the constraints (23)–(25). □

LEMMA 4.3. *Assume (27) and (28). There exists a unique solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ to the fluid model (K, λ, ν) satisfying the nonzero initial condition $(\xi, \mu) \in \mathcal{I}$ on the interval $[0, t^*)$, where either $t^* < \infty$ or $t^* = \infty$. In the case when $t^* < \infty$, the existence and uniqueness can be extended to $[0, t^*]$ with $(\bar{\mathcal{Q}}(t^*), \bar{\mathcal{Z}}(t^*)) = (\mathbf{0}, \mathbf{0})$. In both cases,*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) \neq (\mathbf{0}, \mathbf{0}) \quad \text{for all } t \in [0, t^*).$$

PROOF. Lemma 4.1 establishes the existence and uniqueness on a small interval $[0, t_1']$, where

$$t_1' = \bar{T}(a_1'), \tag{53}$$

$$a_1' = \sup\{u \leq b\colon x(u) > 0\} \tag{54}$$

according to (51) and (52) in the proof of Lemma 4.1, and the constant b is the same as in Lemma A.1 and only depends on ρ and F. We put the subscript one on the quantities corresponding to the first piece. Lemma 4.1 also

says that $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \neq (\mathbf{0}, \mathbf{0})$ on the interval $[0, t_1')$. If $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1')) = (\mathbf{0}, \mathbf{0})$, then let $t^* = t_1'$ and the proof is done and we stop. If $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1')) \neq (\mathbf{0}, \mathbf{0})$, then by (54),

$$a_1' = b. \tag{55}$$

Viewing $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1'))$ as an initial condition, by Lemma 4.1, there exists a unique fluid model solution $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot))$ on the interval $[0, t_2']$, and similar to (53) and (54),

$$t_2' = \bar{T}_2(a_2'),$$
$$a_2' = \sup\{u \le b\colon x_1(u) > 0\},$$

where $\bar{T}_1(\cdot)$ is the corresponding time change based on $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot))$ (defined in the same way as $\bar{T}(\cdot)$ for the process $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$) and $x_1(\cdot)$ is the solution to (50) with $h(\cdot)$ generated by the initial condition $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1'))$ via (48). Again, according to Lemma 4.1, $(\bar{\mathcal{Q}}_1(\cdot), \bar{\mathcal{Z}}_1(\cdot)) \neq (\mathbf{0}, \mathbf{0})$ on the interval $[0, t_2')$. By Lemma 4.2, we obtain a fluid model solution on the interval $[0, t_1' + t_2']$ by defining $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = ((\bar{\mathcal{Q}}_1(t - t_1'), \bar{\mathcal{Z}}_1(t - t_1')))$ for all $t \in (t_1', t_1' + t_2']$. If $(\bar{\mathcal{Q}}(t_1' + t_2'), \bar{\mathcal{Z}}(t_1' + t_2')) = (\mathbf{0}, \mathbf{0})$, then let $t^* = t_1' + t_2'$ and the proof is complete. Otherwise, we have

$$a_2' = b$$

and we can continue the procedure.

If this procedure never stops, then we get a sequence $\{t_i', a_i'\}_{i=1}^{\infty}$ with $a_i' = b$ for all i. Setting

$$t^* = \sum_{i=1}^{\infty} t_i',$$

we have established the existence of a fluid model solution on the interval $[0, t^*)$; the solution never reaches zero before t^*. If $\sum_{i=1}^{\infty} t_i' = \infty$, the proof is complete because the whole interval $[0, \infty)$ is covered. Otherwise, for each $0 \le s < t^*$, there exists an i_s such that $\sum_{i=i_s}^{\infty} t_i' \ge s$. Thus,

$$\lim_{t \to t^*} \bar{S}(s, t) > \sum_{i=i_s}^{\infty} a_i' = \sum_{i=i_s}^{\infty} b = \infty.$$

By the fluid dynamic Equation (22), $\lim_{t \to t^*} \bar{\mathcal{Z}}(t) = \mathbf{0}$. The constraints (24) and (25) imply $\lim_{t \to t^*} \bar{\mathcal{Q}}(t) = \mathbf{0}$. Thus, we can extend the existence of the fluid model solution to the interval $[0, t^*]$ with $(\bar{\mathcal{Z}}(t^*), \bar{\mathcal{Z}}(t^*)) = (\mathbf{0}, \mathbf{0})$. We have now established the existence of fluid model solution. To prove the uniqueness, note that the interval $[0, t^*)$ is covered by $\bigcup_{j=0}^{\infty} [\sum_{i=0}^{j} t_i', \sum_{i=0}^{j+1} t_i']$ (here, we take $t_0 = 0$ for notational convenience). The uniqueness of the solution on the interval $[0, t_1']$ follows directly from Lemma 4.1. The uniqueness on the interval $[t_1', t_1' + t_2']$ can be proved using the same argument in Lemma 4.1 by viewing $(\bar{\mathcal{Q}}(t_1'), \bar{\mathcal{Z}}(t_1'))$ as the initial condition and $(\bar{\mathcal{Q}}(t_1' + \cdot), \bar{\mathcal{Z}}(t_1' + \cdot))$ as the corresponding fluid model solution on the interval $[0, t_2']$. Continuing with this procedure establishes the uniqueness. This completes the proof. □

The following lemma establishes the workload-conserving property for any fluid model solution before it reaches zero.

Lemma 4.4. *Assume* (27) *and* (28). *For the fluid model solution in Lemma* 4.3, *we have the following workload-conserving property on* $[0, t^*)$:

$$\langle \chi, \bar{\mathcal{Q}}(t) \rangle + \langle \chi, \bar{\mathcal{Z}}(t) \rangle = \langle \chi, \xi \rangle + \langle \chi, \mu \rangle + (\rho - 1)t. \tag{56}$$

Proof. By (21) and (22), we have

$$\langle \chi, \bar{\mathcal{Q}}(t) \rangle = \int_0^{\infty} \bar{\mathcal{Q}}(t)(A_y)\, dy = \bar{Q}(t)\beta,$$
$$\langle \chi, \bar{\mathcal{Z}}(t) \rangle = \int_0^{\infty} \mu(A_y + \bar{S}(t))\, dy + \int_0^{\infty} \int_0^{t} \nu(A_y + \bar{S}(s, t))\, d[\lambda s - \bar{Q}(s)]\, dy. \tag{57}$$

Let $\tilde{F}$ be the distribution function associated with the probability measure $(1/\langle 1, \mu\rangle)\mu$, so that $\mu(A_y) = \bar{Z}(0)[1-\tilde{F}(y)]$. Because the cumulative service amount $\bar{S}(\cdot)$ has an inverse on the interval $[0, t^*)$, we can perform the change of variable $u = \bar{S}(t)$ and $t = \bar{T}(u)$ for all $t < t^*$. The first term in (57) becomes

$$\begin{aligned}
&\bar{Z}(0)\int_0^\infty [1-\tilde{F}(y)]\,dy + \bar{Z}(0)\int_0^\infty [\tilde{F}(y)-\tilde{F}(y+u)]\,dy \\
&\quad = \langle \chi, \mu\rangle + \bar{Z}(0)\int_0^u -[1-\tilde{F}(v)]\,dv \\
&\quad = \langle \chi, \mu\rangle - \int_0^u \bar{\mathcal{Z}}(0)(A_v)\,dv
\end{aligned} \tag{58}$$

and, by applying Fubini's theorem, the second term in (57) becomes

$$\begin{aligned}
&\int_0^u \int_0^\infty \nu(A_y + u - v)\,dy\,d[\lambda\bar{T}(v) - \bar{Q}(\bar{T}(v))] \\
&\quad = \beta\int_0^u \int_0^\infty \frac{1-F(y+u-v)}{\beta}\,dy\,d[\lambda\bar{T}(v) - \bar{Q}(\bar{T}(v))] \\
&\quad = \beta\int_0^u [1-F_e(u-v)]\,d[\lambda\bar{T}(v) - \bar{Q}(\bar{T}(v))] \\
&\quad = \lambda\beta\bar{T}(u) - \beta[\bar{Q}(\bar{T}(u)) - \bar{Q}(0)] - \beta\int_0^u F_e(u-v)\,d[\lambda\bar{T}(v) - \bar{Q}(\bar{T}(v))].
\end{aligned} \tag{59}$$

To deal with the last term in the above, perform the change of variable $u = \bar{S}(t)$ and $t = \bar{T}(u)$ for (44). Note that $\bar{T}'(u) = \bar{Z}(\bar{T}(u)) = \bar{Z}(t)$, Thus, we have

$$\begin{aligned}
\bar{T}'(u) &= \mu(A_u) + \beta\int_0^u \frac{1-F(u-v)}{\beta}\,d[\lambda\bar{T}(v) - \bar{Q}(\bar{T}(v))] \\
&= \mu(A_u) + \beta\int_0^u F_e'(u-v)\,d[\lambda\bar{T}(v) - \bar{Q}(\bar{T}(v))].
\end{aligned}$$

Integrating both sides of the above equation yields

$$\bar{T}(u) = \int_0^u \mu(A_v)\,dv + \beta\int_0^u F_e(u-v)\,d[\lambda\bar{T}(v) - \bar{Q}(\bar{T}(v))]. \tag{60}$$

The proof is completed by combining (58), (59), and (60) and substituting $\bar{T}(u)$ with t. □

4.2. Starting with zero initial condition when $\rho \le 1$. Intuitively, the fluid model solution should stay at zero forever in this case. We rigorously prove this result in the following lemma.

LEMMA 4.5. *When $\rho \le 1$, $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv (\mathbf{0}, \mathbf{0})$ is the unique solution to the fluid model (λ, K, ν) with initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$.*

PROOF. Note that $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv \mathbf{0}$ implies $\bar{Z}(\cdot) \equiv 0$. By (20), we have $\bar{S}(s,t) = \infty$ for all $t > s \ge 0$. Because ν is a measure on $\mathbb{R}$, there is no mass at infinity by definition so $\nu(A_y + \bar{S}(s,t)) = 0$ for all $y \ge 0$. This implies that the integral on the right-hand side of (22) is zero, so $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv \mathbf{0}$ satisfies Equation (22). It is clear that fluid dynamic Equation (21) and constraints (23) through (25) are satisfied, so $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \equiv \mathbf{0}$ is a fluid model solution.

We now prove that it is the only solution. If $(\mathbf{0}, \mathbf{0})$ is the unique fluid model solution on the interval $[0, K/\lambda]$, then, by Lemma 4.2, we can extend the uniqueness to $[K/\lambda, 2K/\lambda]$ and so on to $[0, \infty)$. Otherwise, there is another solution on $[0, K/\lambda]$, which is denoted by $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$. By (21) and (22), for any fluid model solution $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$ starting at $(\mathbf{0}, \mathbf{0})$, we have

$$\begin{aligned}
\bar{X}^\dagger(t) &= \bar{\mathcal{Q}}^\dagger([0,\infty)) + \bar{\mathcal{Z}}^\dagger((0,\infty)) \\
&\le Q^\dagger(t) + \int_0^t 1\,d[\lambda s - \bar{Q}^\dagger(s)] \le \lambda t.
\end{aligned}$$

Thus, $\bar{Q}^\dagger(\cdot) \equiv 0$ on the interval $[0, K/\lambda]$ by (24) and by (22), the workload process satisfies for all $t \geq 0$,

$$\begin{aligned}\bar{W}^\dagger(t) &= \int_0^\infty \int_0^t \nu(A_y + \bar{S}^\dagger(s,t))\, d\lambda s\, dy \\ &= \lambda \int_0^t \int_0^\infty \nu(A_y + \bar{S}^\dagger(s,t))\, dy\, ds,\end{aligned}$$

where the second equality is because of Fubini's theorem. Because

$$\int_0^\infty \nu(A_y + \bar{S}^\dagger(s,t))\, dy \leq \int_0^\infty \nu(A_y)\, dy < \infty,$$

$\bar{W}^\dagger(\cdot)$ is continuous on $[0, K/\lambda]$. Note that

$$\bar{W}^\dagger(0) = 0 \tag{61}$$

but it is different from $(\mathbf{0}, \mathbf{0})$, so there must be a $t_1 \in (0, K/\lambda]$ such that $(\bar{\mathcal{Q}}^\dagger(t_1), \bar{\mathcal{Z}}^\dagger(t_1)) \neq (\mathbf{0}, \mathbf{0})$, which implies that

$$\bar{W}^\dagger(t_1) > 0. \tag{62}$$

Let $t_0 = \sup_{0 \leq t < t_1} \{\bar{W}^\dagger(t) = 0\}$, then $0 \leq t_0 < t_1$ by (61) and (62) and continuity of $\bar{W}^\dagger(\cdot)$. Again, by continuity of $\bar{W}^\dagger(\cdot)$, there exists a $t_\delta \in (t_0, t_1)$ such that $\bar{W}^\dagger(t_\delta) = \delta < \bar{W}^\dagger(t_1)$ for some $\delta > 0$. On the interval $[t_\delta, t_1]$, $(\bar{\mathcal{Q}}^\dagger(\cdot), \bar{\mathcal{Z}}^\dagger(\cdot))$ never reaches zero. Thus, by Lemmas 4.2 and 4.4,

$$\bar{W}^\dagger(t_\delta + t) = \bar{W}^\dagger(t_\delta) + (1-\rho)t \quad \text{for } t \in [0, t_1 - t_\delta].$$

This implies that $\bar{W}^\dagger(t_1) \leq \bar{W}^\dagger(t_\delta)$, which is a contradiction. □

4.3. Starting with zero initial condition when $\rho > 1$. In Jean-Marie and Robert [18] and Puha et al. [26], a very nice approach has been developed for overloaded PS queue with zero initial condition. We can apply the same approach to the LPS queue without much adjustment because the fluid models of the LPS queue and PS queue behave the same until the time that total job size becomes larger than K.

Intuitively, the fluid model solution should grow as time goes by. Let us first assume that the fluid queue length process $\bar{X}(\cdot)$ grows linearly on a small interval, i.e.,

$$\begin{aligned}\bar{Q}(t) &= \langle 1, \bar{\mathcal{Q}}(t) \rangle = 0, \\ \bar{Z}(t) &= \langle 1, \bar{\mathcal{Z}}(t) \rangle = mt,\end{aligned} \tag{63}$$

for all $t \in [0, K/m]$, where $m > 0$ is to be determined. The following analysis is taken from Puha et al. [26]. By (25),

$$\bar{S}(s,t) = \int_s^t \frac{1}{m\tau}\, d\tau = \frac{1}{m} \log \frac{t}{s}, \quad 0 < s < t \leq K/m. \tag{64}$$

Plug (63) and (64) into (44) to get

$$mt = \lambda \int_0^t \left[1 - F\left(\frac{1}{m} \log \frac{t}{s}\right)\right] ds \quad \text{for all } t \leq \frac{K}{m}.$$

Perform the change of variable $v = (1/m)\log(t/s)$ to obtain

$$\frac{1}{\lambda\beta} mt = mt \int_0^\infty e^{-mv} \frac{1 - F(v)}{\beta}\, dv \quad \text{for all } t \leq \frac{K}{m}.$$

By the definition of F_e and ρ, we must have

$$\int_0^\infty e^{-mv}\, dF_e(v) = \frac{1}{\rho}. \tag{65}$$

Note that the left-hand side is the Laplace transform of the distribution F_e. As a function of $m \in (0, \infty)$, it is strictly decreasing and maps onto $(0, 1)$. Because $\rho > 1$ in this case, (65) has a unique solution, which we denote by m_ρ^*. Now, let

$$\begin{aligned}\bar{\mathcal{Q}}(t)(A_y) &= 0, \\ \bar{\mathcal{Z}}(t)(A_y) &= \lambda \int_0^t \left[1 - F\left(y + \frac{1}{m_\rho^*} \log \frac{t}{s}\right)\right] ds\end{aligned} \tag{66}$$

for all $t \in [0, K/m_\rho^*]$ and $y \geq 0$. It is clear that $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ is a fluid model solution on the interval $[0, K/m_\rho^*]$. By Lemma 4.2, $(\bar{\mathcal{Q}}(K/m_\rho^* + \cdot), \bar{\mathcal{Z}}(K/m_\rho^* + \cdot))$ can be viewed as the fluid model solution with initial condition $(\bar{\mathcal{Q}}(K/m_\rho^*), \bar{\mathcal{Z}}(K/m_\rho^*))$, which exists on $[0, \infty)$. Thus, we have found a fluid model solution with zero initial condition.

Similarly, as in the case $\rho \leq 1$, the difficulty is to prove uniqueness. Puha et al. [26] has established existence and uniqueness of fluid model solutions for overloaded PS queue with zero initial condition. We can borrow the result for the reason that the total fluid amount of jobs of any fluid model solution starting at zero is bounded by λt at any time $t \geq 0$, as explained in the proof of Lemma 4.5. The sharing limit K is never reached on the interval $[0, K/\lambda]$, so the model is the same as a standard PS queue. In fact, the fluid dynamic Equation (22) is what is needed in Theorem 4.2 and Lemma 4.10, which implies uniqueness on the interval $[0, K/\lambda]$. The uniqueness can be extended to $[K/\lambda, \infty)$ by Lemma 4.2 because $\bar{X}(K/\lambda) > 0$. Thus, we have the following result.

LEMMA 4.6. *Assume* (27) *and* (28). *When* $\rho > 1$, *there exists a unique solution to the fluid model* (K, λ, ν) *with initial condition* $(\mathbf{0}, \mathbf{0})$.

We are now in a position to sum up all the above cases and prove all results on the fluid model.

PROOF OF THEOREM 3.1. If the initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$, then the result is established by Lemmas 4.5 and 4.6. If $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$, then, by Lemma 4.3, either we have existence and uniqueness on the interval $[0, \infty)$ and the proof is done or the result holds on a finite interval $[0, t^*]$ with $(\bar{\mathcal{Q}}(t^*), \bar{\mathcal{Z}}(t^*)) = (\mathbf{0}, \mathbf{0})$. The result is then established by applying Lemmas 4.2 and 4.5. □

PROOF OF PROPOSITION 3.1. If $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$ and $\rho \leq 1$, then it follows from Lemma 4.5 that $\bar{W}(t) = (0 + (\rho - 1)t)^+$. If $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$ and $\rho > 1$, for any $t \in [0, K/m_\rho^*]$, take the integration of both sides of (66) with respect to y to get

$$\bar{W}(t) = \langle \chi, \bar{\mathcal{Z}}(t) \rangle = \lambda \int_0^\infty \int_0^t \left[1 - F\left(y + \frac{1}{m_\rho^*} \log \frac{t}{s} \right) \right] ds\, dy.$$

Performing the change of variable $v = (1/m_\rho^*) \log(t/s)$ and applying Fubini's theorem, we obtain $\bar{W}(t) = 0 + (\rho - 1)t$. If $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$, then the workload-conserving property holds before the fluid model solution reaches zero. Note that the fluid model solution reaches zero if and only if the workload reaches zero. So, when $\rho \geq 1$, $\bar{W}(t) = w + (\rho - 1)t > 0$ for all $t > 0$ and the result holds on $[0, \infty)$. When $\rho < 1$, $\bar{W}(t_w) = 0$ for $t_w = w/(1 - \rho)$. By weak stability, $\bar{W}(t) = 0$ for all $t \geq t_w$. □

PROOF OF THEOREM 3.2. Weak stability is already proved in Lemma 4.5. Because the descriptor $(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t))$ equals $(\mathbf{0}, \mathbf{0})$ if and only if $\bar{W}(t) = 0$, the stability follows immediately from Proposition 3.1. □

5. Precompactness. The objective of this section is to show the precompactness property (Theorem 5.1) for the fluid-scaled processes $(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot))$ defined in §3.2.

Consider the rth system. A fluid-scaled version of stochastic dynamic Equations (7) and (8) can be written as

$$\bar{\mathcal{Q}}^r(t)(A') = \frac{1}{r} \sum_{i=r\bar{B}^r(t)+1}^{r\bar{E}^r(t)} \delta_{v_i^r}(A'),$$

$$\bar{\mathcal{Z}}^r(t)(A) = \frac{1}{r} \sum_{i=-r\bar{X}^r(0)+1}^{-r\bar{Q}^r(0)} \delta_{\tilde{v}_i^r}(A + \bar{S}^r(t)) + \frac{1}{r} \sum_{i=-r\bar{Q}^r(0)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}(A + \bar{S}^r(\tau_i, t))$$

for $t \geq 0$ and any Borel sets $A' \subseteq [0, \infty)$ and $A \subseteq (0, \infty)$. Thus, by the above equations, we have for $0 \leq s \leq t$

$$\bar{\mathcal{Q}}^r(t)(A') = \bar{\mathcal{Q}}^r(s)(A') + \frac{1}{r} \sum_{i=r\bar{E}^r(s)+1}^{r\bar{E}^r(t)} \delta_{v_i^r}(A') - \frac{1}{r} \sum_{i=r\bar{B}^r(s)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}(A'), \tag{67}$$

$$\bar{\mathcal{Z}}^r(t)(A) = \bar{\mathcal{Z}}^r(s)(A + \bar{S}^r(s, t)) + \frac{1}{r} \sum_{i=r\bar{B}^r(s)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}(A + \bar{S}^r(\tau_i, t)). \tag{68}$$

The dynamics of the system is determined by the above equations. Equation (67) says that the status of the buffer at time t equals the status at time s plus what has arrived to the buffer and minus what has left from

the buffer during time interval $(s, t]$. Those jobs that left the buffer enter service, and the service process has been taken care of by shifting the set A by the cumulative service amount $S^r(\tau_i, t)$ that the ith job receives. This corresponds to the second term on the right-hand side of (68). This plus the status at time s shifted by accumulative service amount $S^r(s, t)$ is equal to the status of the server at time t, as indicated in (68). To simplify the notation in this section, for all $0 \le s \le t$, denote

$$\bar{E}^r(s, t) = \bar{E}^r(t) - \bar{E}^r(s), \qquad \bar{B}^r(s, t) = \bar{B}^r(t) - \bar{B}^r(s).$$

Note that $\bar{\mathcal{Z}}^r(t) \in \mathbf{M}_2$ on each sample path for each $r > 0$ and $t > 0$. Because of the convention that $\mathbf{M}_2$ can be embedded in $\mathbf{M}_1$ (cf. §1.1), we view $\bar{\mathcal{Z}}^r(t)$ as an element in $\mathbf{M}_1$ when it is convenient. In particular, $\bar{\mathcal{Z}}^r(t)(A)$ is well-defined for each Borel set $A \subset [0, \infty)$.

The compact containment property is derived in §5.1. Section 5.2 serves as a preparation for the oscillation bound. The oscillation bound is then proved in §5.3, followed by the precompactness result (Theorem 5.1). The framework of the proofs is similar to that of Gromoll and Kruk [13] and Gromoll et al. [15].

5.1. Compact containment. The main objective of this section is to establish the compact containment property in Lemma 5.4, which is the first main step to prove precompactness. First, let us establish a bound for the arrival processes.

Fix $T > 0$. It follows immediately from condition (36) that for each ϵ, $\epsilon' > 0$, there exists an r_0 such that when $r > r_0$,

$$\mathbb{P}^r\left(\sup_{0 \le s < t \le T} |\bar{E}^r(s, t) - \lambda(t - s)| < \epsilon'\right) \ge 1 - \epsilon. \tag{69}$$

To facilitate some arguments later on, we derive the following result from the above inequality.

Lemma 5.1. *Fix $T > 0$. There exists a function $\epsilon_E(\cdot)$, which vanishes at infinity such that*

$$\mathbb{P}^r\left(\sup_{0 \le s < t \le T} |\bar{E}^r(s, t) - \lambda(t - s)| < \epsilon_E(r)\right) \ge 1 - \epsilon_E(r)$$

for each $r \ge 0$.

Proof. For each index r, let

$$H_r = \{\delta > 0\text{: (69) is true for } \epsilon' = \epsilon = \delta\}.$$

Clearly H_r is not empty because $1 \in H_r$. Let $\epsilon_E(r) = \inf H_r$ for each $r \ge 0$. Assume that $\epsilon_E(r)$ does not vanish at infinity. There exists a $\delta > 0$ and a subsequence $\{r_n\}_{n=1}^{\infty}$, which increases to infinity such that

$$\epsilon_E(r_n) \ge \delta \quad \text{for all } n \ge 0. \tag{70}$$

However, for $\epsilon' = \epsilon = \delta/2$, there exists an r_δ such that when $r_n \ge r_\delta$, (69) must hold. This contradicts (70). □

Denote

$$\Omega_E^r = \left\{\sup_{t \in [0, T]} |\bar{E}^r(t) - \lambda t| < \epsilon_E(r)\right\}.$$

We have that

$$\lim_{r \to \infty} \mathbb{P}^r(\Omega_E^r) = 1. \tag{71}$$

It is clear from the policy constraint (10) that for all $t \ge 0$,

$$\bar{Z}^r(t) \le K^r/r < K + 1, \tag{72}$$

where the last inequality holds for all large r because $K^r/r \to K$. Lemma 5.2 establishes a bound for the buffer size $\bar{Q}^r(\cdot)$.

Lemma 5.2. *Assume (36) and (41). Fix $T > 0$. For each $\eta > 0$, there exists a constant $M_1 > 0$ such that*

$$\liminf_{r \to \infty} \mathbb{P}^r\left(\sup_{t \in [0, T]} \bar{Q}^r(t) < M_1\right) \ge 1 - \eta.$$

PROOF. Plugging $A=[0,\infty)$ in (67) and letting $s=0$, we get

$$\bar{Q}^r(t) \le \bar{Q}^r(0) + \bar{E}^r(t). \tag{73}$$

By condition (41), there exists a constant M' such that

$$\liminf_{r\to\infty} \mathbb{P}^r(\bar{Q}^r(0) < M') \ge 1-\eta.$$

By (71) and (73), we have that

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\sup_{t\in[0,T]} \bar{Q}^r(t) < M' + \lambda T + 1\right) \ge 1-\eta.$$

Lemma 5.2 is proved by letting $M_1 = M' + \lambda T + 1$. □

LEMMA 5.3. *Assume* (36)–(42). *Fix* $T>0$. *For any* $\eta>0$, *there exists a constant* $M_2>0$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\sup_{t\in[0,T]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t)\rangle < M_2\right) > 1-\eta,$$

where the positive constant p *is the same as in conditions* (38) *and* (42).

PROOF. By condition (42),

$$\liminf_{r\to\infty} \mathbb{P}^r(\langle \chi^{1+p}, \bar{\mathcal{Z}}^r(0)\rangle < \langle \chi^{1+p}, \xi^* + \mu^*\rangle + 1) = 1.$$

Denote the event in the above by Ω_0^r. By Lemma 5.2, for any $\eta>0$, there exists a constant $M_1>0$ such that

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\sup_{t\in[0,T]} \bar{Q}^r(t) < M_1\right) > 1-\eta.$$

Denote the event in the above by $\Omega_1^r(M_1)$. Note that on the event $\Omega_1^r(M_1)\cap\Omega_E^r$,

$$\langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t)\rangle \le \langle \chi^{1+p}, \bar{\mathcal{Z}}^r(0)\rangle + \frac{1}{r}\sum_{i=-rM_1}^{\lfloor \lambda r T + r\epsilon_E(r)\rfloor} \langle \chi^{1+p}, \delta_{v_i^r}\rangle \tag{74}$$

for any $t\in[0,T]$. By condition (38), $\langle \chi^{1+p}, \nu^r\rangle < \infty$ and $\langle \chi^{1+p}, \nu\rangle < \infty$. Because we only need to consider large enough r such that $\epsilon_E(r)<1$, by Lemma A.2 in Gromoll et al. [14],

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\frac{1}{r}\sum_{i=-rM_1}^{\lfloor \lambda r T + r\epsilon_E(r)\rfloor} \langle \chi^{1+p}, \delta_{v_i^r}\rangle < (\lambda T + M_1 + 1)\langle \chi^{1+p}, \nu\rangle + 1\right) = 1.$$

Denote the above event by $\Omega_p^r(M_1)$. Then, by (71), we have

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_E^r \cap \Omega_0^r \cap \Omega_1^r(M_1) \cap \Omega_p^r(M_1)) > 1-\eta. \tag{75}$$

Lemma 5.3 is proved by letting $M_2 = \langle \chi^{1+p}, \xi^* + \mu^*\rangle + (\lambda T + M_1 + 1)\langle \chi^{1+p}, \nu\rangle + 2$. □

Denote

$$\Omega_B^r(M) = \left\{\sup_{t\in[0,T]} \bar{Q}^r(t) < M \text{ and } \sup_{t\in[0,T]} \bar{Z}^r(t) < M\right\} \cap \left\{\sup_{t\in[0,T]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t) + \bar{\mathcal{Z}}^r(t)\rangle < M\right\}.$$

By (72) and Lemmas 5.2 and 5.3, for any $\eta>0$, there exists a constant $M > K+1$ such that

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_B^r(M)) > 1-\eta. \tag{76}$$

A set $\mathbf{K}\subset\mathbf{M}_1$ is relatively compact if $\sup_{\xi\in\mathbf{K}} \xi(\mathbb{R}_+) < \infty$ and if there exists a sequence of nested compact sets $J_n \subset \mathbb{R}_+$ such that $\cup J_n = \mathbb{R}_+$ and

$$\lim_{n\to\infty} \sup_{\xi\in\mathbf{K}} \xi(J_n^c) = 0,$$

where J_n^c denotes the complement of J_n; see Kallenberg [19], Theorem A7.5. Denote

$$\mathbf{K}(M) = \{\xi\in\mathbf{M}_1 \colon \xi(\mathbb{R}_+) < M \text{ and } \xi((n,\infty)) \le M/n \text{ for all } n\in\mathbb{Z}_+\}.$$

Clearly, $\mathbf{K}(M)$ is a relatively compact set for any constant $M>0$.

LEMMA 5.4. *On the event* $\Omega_B^r(M)$,

$$\bar{\mathcal{Q}}^r(t) \in \mathbf{K}(M) \qquad \text{and} \qquad \bar{\mathcal{Z}}^r(t) \in \mathbf{K}(M) \quad \text{for all } t \in [0, T]$$

PROOF. Note that both $\sup_{t\in[0,T]} \bar{\mathcal{Q}}^r(t)([0,\infty))$ and $\sup_{t\in[0,T]} \bar{\mathcal{Z}}^r(t)((0,\infty))$ are bounded by M according to the definition of $\Omega_B^r(M)$. By Markov's inequality, for any $t \geq 0$,

$$\bar{\mathcal{Q}}^r(t)((n,\infty)) \leq \frac{\langle \chi^{1+p}, \bar{\mathcal{Q}}^r(t)\rangle}{n^{1+p}},$$

which is bounded by M/n^{1+p} by the definition of $\Omega_B^r(M)$. The same argument applies for $\bar{\mathcal{Z}}^r(t)$. □

5.2. Asymptotic regularity. The second major step to prove precompactness is to obtain the oscillation bound in §5.3. Oscillations mainly result from sudden departures of a large number of jobs. To control the departure process, we show that $\bar{\mathcal{Z}}^r(\cdot)$ assigns arbitrarily small mass to small intervals. Similar results have been proved for PS queues and related models: see Gromoll and Kruk [13] and Gromoll et al. [15]. In our model, the process of jobs entering the server is $\bar{B}^r(t) = \bar{E}^r(t) - \bar{Q}^r(t)$ instead of $\bar{E}^r(t)$, which creates additional difficulties.

Note the Glivenko-Cantelli estimate in Lemma B.1. By the same argument as in Lemma 5.1, for fixed M, $T > 0$, there exists a function $\epsilon_{\mathrm{GC}}(\cdot)$, which vanishes at infinity such that the probability inequality in Lemma B.1 holds with ϵ and ϵ' replaced by this function. In other words, denote

$$\Omega_{\mathrm{GC}}^r(M) = \left\{ \max_{-rM<n<r(M+2\lambda T)} \sup_{l\in[0,\,2M+2\lambda T]} \sup_{f\in\bar{\mathcal{V}}} |\langle f, \bar{\eta}^r(n,l)\rangle - l\langle f, \nu^r\rangle| < \epsilon_{\mathrm{GC}}(r) \right\},$$

where

$$\bar{\eta}^r(n,l) = \frac{1}{r}\sum_{i=n+1}^{n+\lfloor rl\rfloor} \delta_{v_i^r}$$

and $\bar{\mathcal{V}}$ is a set of functions of the form $1_{(x,\infty)}$ and $1_{[x,\infty)}$ for all $x \in \mathbb{R}_+$ with an envelope function $\bar{f}$ (see Appendix B).

We have

$$\lim_{r\to\infty} \mathbb{P}^r(\Omega_{\mathrm{GC}}^r(M)) = 1. \tag{77}$$

The Glivenko-Cantelli estimate helps prove the following result.

LEMMA 5.5. *Assume* (36)–(43). *Fix* $T > 0$. *For each* $\epsilon, \eta > 0$, *there exists a* $\kappa > 0$ (*depending on* ϵ *and* η) *such that*

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^r(t)([x, x+\kappa]) \leq \epsilon \right) \geq 1 - \eta. \tag{78}$$

PROOF. We first show that for any $\epsilon, \eta > 0$, there exists a κ such that

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^r(0)([x, x+\kappa]) \leq \epsilon/2 \right) \geq 1 - \eta/2. \tag{79}$$

It follows from the initial condition (41) that $\bar{\mathcal{Z}}^r(0) \Rightarrow \mu^*$ as $r \to \infty$. Because μ^* is a finite Borel measure, there exists an $M > 0$ such that

$$\mu^*([M,\infty)) < \epsilon/4.$$

By (43), the distribution function associated with the measure μ^* is continuous and is thus uniformly continuous on the finite interval $[0, 2M]$. Hence, there exists a $\kappa \in (0, M]$ such that

$$\sup_{x\in[0,\,M]} \mu^*([x, x+\kappa]) < \epsilon/4.$$

The above two inequalities imply

$$\sup_{x\in\mathbb{R}_+} \mu^*([x, x+\kappa]) < \epsilon/4.$$

Let $N = \lceil M/\kappa \rceil$. Denote $I_n = [n\kappa, (n+2)\kappa]$ for $n = 0, 1, \ldots, N-1$, and $I_N = [M, \infty)$. Note that for every $x \in [0, \infty)$, there exists an $n \le N$ such that $[x, x+\kappa] \subset I_n$. To prove (79), it suffices to show

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\sup_{n\le N} \bar{\mathcal{Z}}^r(0)(I_n) \le \epsilon/2\right) \ge 1 - \eta/2. \tag{80}$$

Denote $\mathbf{A} = \{\mu \in \mathbf{M}_2\colon \max_{n\le N} \mu(I_n) < \epsilon/2\}$. It is clear that $\mu^* \in \mathbf{A}$. Now, let us prove that the set $\mathbf{A}$ is open in the space $\mathbf{M}_2$ equipped with the Prohorov metric. Let $\{\mu_k\} \subset \mathbf{M}_2$ be a sequence in the Polish space $\mathbf{M}_2$, satisfying $\mu_k \to \mu$ for some $\mu \in \mathbf{A}$. Because each I_n is closed, by the Portmanteau theorem (Theorem 2.1 in Billingsley [3] adapted to finite measures; see also Gromoll et al. [15]),

$$\limsup_{k\to\infty} \mu_k(I_n) \le \mu(I_n) < \epsilon/2 \quad \text{for all } n \le N.$$

Hence, $\mu_k \in \mathbf{A}$ for all sufficiently large k, which implies that $\mathbf{A}$ is open in $\mathbf{M}$. Thus, a second application of the Portmanteau theorem yields

$$\liminf_{r\to\infty} \mathbb{P}^r(\bar{Z}^r(0) \in \mathbf{A}) \ge \mathbb{P}(\mu^* \in \mathbf{A}) = 1,$$

which implies (80).

Now, we need to extend this result to the interval $[0, T]$. Denote the event in (79) by Ω_1^r. Let

$$\Omega_2^r(M) = \Omega_1^r \cap \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(M).$$

By (71), (76), and (77), there exists an $M > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_2^r(M)) \ge 1 - \eta.$$

In the remainder of the proof, all random objects are evaluated at a fixed sample path in $\Omega_2^r(M)$.

For any $r > 0$, $t \in [0, T]$, we define the random time

$$t_0 = \sup\{\{s \le t\colon \langle 1, \bar{\mathcal{Z}}^r(s)\rangle < \epsilon/4\} \cup \{0\}\}.$$

If $t_0 = 0$, then by (79), for each $x \in \mathbb{R}_+$,

$$\bar{\mathcal{Z}}^r(0)([x, x+\kappa] + \bar{S}^r(t)) \le \epsilon/2.$$

If $t_0 \in (0, t]$, then for each $\delta > 0$, there exists an s such that $t_0 - \delta < s < t_0$ and $\bar{Z}^r(s)(\mathbb{R}_+) < \epsilon/4$. Because we are only concerned with small ϵ (which should be small enough such that $\bar{Z}^r(s) < \epsilon/4 < K^r/r$), $\bar{Q}^r(s) = 0$ by the policy constraint (10). Note that (5) implies

$$\bar{B}^r(s', t) \le \bar{E}^r(s', t) + \bar{Q}^r(s') \quad \text{for all } s' \le t. \tag{81}$$

Because we are on the event Ω_E^r, for any $\epsilon_1 > 0$, we have $\bar{B}^r(s, t_0) \le \lambda\delta + \epsilon_1$ for all large enough r. For any Borel set A, by the fluid-scaled system dynamic Equation (68),

$$\bar{\mathcal{Z}}^r(t_0)(A) \le \bar{\mathcal{Z}}^r(s)(\mathbb{R}_+) + \bar{B}^r(s, t_0) \le \epsilon/4 + \lambda\delta + \epsilon_1,$$

which can be made smaller than $\epsilon/2$ by choosing ϵ_1, δ suitably small.

The fluid-scaled stochastic dynamic equation over the interval $[t_0, t]$ can be written as

$$\bar{\mathcal{Z}}^r(t)([x, x+\kappa]) = \bar{\mathcal{Z}}^r(t_0)([x, x+\kappa] + \bar{S}^r(t_0, t)) + \frac{1}{r} \sum_{i=r\bar{B}^r(t_0)+1}^{r\bar{B}^r(t)} \delta_{v_i^r}([x, x+\kappa] + \bar{S}^r(\tau_i, t))$$

for each $x \in \mathbb{R}_+$. By the choice of t_0, the first term on the right-hand side of the above equation is always upper bounded by $\epsilon/2$. Let I denote the second term on the right-hand side of the above equation. Now, it only remains to show that $I < \epsilon/2$.

Let $t_0, t_1, \ldots, t_N = t$ be a partition of the interval $[t_0, t]$ such that $|t_{j+1} - t_j| < \delta$ for all $j = 0, \ldots, N-1$, where δ and N are to be chosen below. Write I as the summation

$$I = \sum_{j=0}^{N-1} \frac{1}{r} \sum_{i=r\bar{B}^r(t_j)+1}^{r\bar{B}^r(t_{j+1})} \delta_{v_i^r}([x, x+\kappa] + \bar{S}^r(\tau_i, t)).$$

Recall that τ_i^r is the time that the ith job starts service, so on each subinterval $[t_j, t_{j+1}]$, the i's to be summed must satisfy $t_j \le \tau_i^r \le t_{j+1}$. This implies that

$$\bar{S}^r(t_{j+1}, t) \le \bar{S}^r(\tau_i, t) \le \bar{S}^r(t_j, t).$$

By the definition of t_0, we have $\bar{Z}^r(s) \ge \epsilon/4$ for all $s \in [t_0, t]$. Thus,

$$\bar{S}^r(t_j, t_{j+1}) \le \frac{4\delta}{\epsilon}.$$

Let

$$C_j = \left[x + \bar{S}^r(t_{j+1}, t), x + \bar{S}^r(t_{j+1}, t) + \kappa + \frac{4\delta}{\epsilon}\right].$$

Then,

$$I \le \sum_{j=0}^{N-1} \frac{1}{r} \sum_{i=r\bar{B}^r(t_j)+1}^{r\bar{B}^r(t_{j+1})} \delta_{v_i^r}(C_j).$$

Because we are on the event $\Omega_E^r \cap \Omega_B^r(M)$, by (81), we have for all $j = 0, \ldots, N-1$:

$$-rM \le r\bar{B}^r(t_j) \le r(\lambda T + \epsilon_1 + M) \le 2\lambda r T + rM,$$

$$\bar{B}^r(t_j, t_{j+1}) \le \lambda T + \epsilon_1 + M \le 2\lambda T + M.$$

Because we are on the event $\Omega_{\mathrm{GC}}^r(M)$,

$$\left| \frac{1}{r} \sum_{i=r\bar{B}^r(t_j)+1}^{r\bar{B}^r(t_{j+1})} \delta_{v_i^r}(C_j) - (\bar{B}^r(t_{j+1}) - \bar{B}^r(t_j))\nu^r(C_j) \right| < \epsilon_1.$$

Thus,

$$I \le \sum_{j=0}^{N-1} [\bar{B}^r(t_{j+1}) - \bar{B}^r(t_j)]\nu^r(C_j) + N\epsilon_1.$$

By (37), for all $\epsilon_2 > 0$,

$$d[\nu^r, \nu] \le \epsilon_2$$

for all large enough r. Note that C_j is a closed Borel set. By the definition of Prohorov metric, we have

$$\nu^r(C_j) \le \nu(C_j^{\epsilon_2}) + \epsilon_2$$

for all large enough r. Because $C_j^{\epsilon_2}$ is a closed interval with length $\kappa + 4\delta/\epsilon + 2\epsilon_2$, by (39), we can choose κ, δ, ϵ_2 small enough such that

$$\nu(C_j^{\epsilon_2}) < \frac{\epsilon}{4(2\lambda T + M)}.$$

Thus, we conclude that

$$\begin{aligned} I &\le \left(\epsilon_2 + \frac{\epsilon}{4(2\lambda T + M)}\right) \sum_{j=0}^{N-1} [\bar{B}^r(t_{j+1}) - \bar{B}^r(t_j)] + N\epsilon_1 \\ &\le \left(\epsilon_2 + \frac{\epsilon}{4(2\lambda T + M)}\right) [\bar{B}^r(t) - \bar{B}^r(t_0)] + N\epsilon_1 \\ &\le \epsilon_2(2\lambda T + M) + \epsilon/4 + N\epsilon_1, \end{aligned}$$

where the last inequality is because we are on the event $\Omega_E^r \cap \Omega_B^r(M)$. Finally, by choosing ϵ_1, ϵ_2 small enough, we obtain that $I < \epsilon/2$. □

In addition to the asymptotic regularity for the server $\bar{\mathcal{X}}^r(\cdot)$, we also have the same property for the buffer $\bar{\mathcal{Q}}^r(\cdot)$. The proof is much easier.

Lemma 5.6. *Assume* (36)–(43). *Fix* $T > 0$. *For each* ϵ, $\eta > 0$, *there exists a* $\kappa > 0$ (*depending on* ϵ *and* η) *such that*

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Q}}^r(t)([x, x+\kappa]) \le \epsilon\right) \ge 1 - \eta. \tag{82}$$

PROOF. Let

$$\Omega_3^r(M) = \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(M).$$

By (71), (76), and (77), there exists an $M > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_3^r(M)) \geq 1 - \eta.$$

In the remainder of the proof, all random objects are evaluated at a fixed sample path in $\Omega_3^r(M)$.

Because we are on the event $\Omega_E^r \cap \Omega_B^r(M)$, $|\bar{B}^r(\cdot)|$ and $\bar{E}^r(\cdot)$ are bounded above by $M + 2\lambda T$. Because we are on the event $\Omega_{\mathrm{GC}}^r(M)$, for any $t \in [0, T]$ and $\epsilon_1 > 0$,

$$\begin{aligned}
&|\bar{\mathcal{Q}}^r(t)([x, x+\kappa]) - (\bar{E}^r(t) - \bar{B}^r(t))\nu^r([x, x+\kappa])| \\
&\quad = \left| \frac{1}{r} \sum_{i=r\bar{B}^r(t)+1}^{r\bar{E}^r(t)} \delta_{v_i^r}([x, x+\kappa]) - (\bar{E}^r(t) - \bar{B}^r(t))\nu^r([x, x+\kappa]) \right| \\
&\quad \leq \epsilon_1
\end{aligned}$$

for all large r. Thus,

$$\begin{aligned}
\bar{\mathcal{Q}}^r(t)([x, x+\kappa]) &\leq (\bar{E}^r(t) - \bar{B}^r(t))\nu^r([x, x+\kappa]) + \epsilon_1 \\
&\leq 2M\nu^r([x, x+\kappa]) + \epsilon_1
\end{aligned}$$

for all large r. By (37), for any $\epsilon_2 > 0$,

$$d[\nu^r, \nu] \leq \epsilon_2$$

for all large enough r. By the definition of Prohorov metric, we have

$$\nu^r([x, x+\kappa]) \leq \nu([x - \epsilon_2, x + \kappa + \epsilon_2]) + \epsilon_2$$

for all large enough r. By (39), we can choose κ, ϵ_2 small enough such that

$$\nu([x - \epsilon_2, x + \kappa + \epsilon_2]) < \epsilon_1.$$

Thus, we conclude that for any $t \in [0, T]$,

$$\bar{\mathcal{Q}}^r(t)[x, x+\kappa] \leq 2M(\epsilon_1 + \epsilon_2) + \epsilon_1.$$

The proof is completed by choosing ϵ_1 and ϵ_2 to be less than $\epsilon/8M$. □

5.3. Oscillation bound. In this section, we use the regularity result in Lemma 5.5 to obtain the oscillation bound in Lemma 5.7. The proof technique of Lemma 5.7 is a simplification of that for Lemma 4.14 in Gromoll and Kruk [13]. Consider a *càdlàg* function $\zeta(\cdot)$ on a fixed interval $[0, T]$ taking values in a metric space $(\mathbf{E}, \pi)$. For $T \geq 0$ and $\delta > 0$, define the *modulus of continuity* to be

$$\mathbf{w}_T(\zeta(\cdot), \delta) = \sup_{s,\, t\in[0,\, T],\, |s-t|<\delta} \pi[\zeta(s), \zeta(t)].$$

If the metric space is $\mathbb{R}$, we just use the Euclidean metric; if the space is $\mathbf{M}_1 \times \mathbf{M}_2$, we use the Prohorov metric $\mathbf{d}$ defined in §1.

LEMMA 5.7. *Assume* (36)–(43). *Fix* $T > 0$. *For each* ϵ, $\eta > 0$, *there exists a* $\delta > 0$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}^r(\max(\mathbf{w}_T(\bar{\mathcal{Q}}^r(\cdot), \delta), \mathbf{w}_T(\bar{\mathcal{Z}}^r(\cdot), \delta)) \leq \epsilon) \geq 1 - \eta. \tag{83}$$

PROOF. For any κ, $\epsilon > 0$, define

$$\Omega_{\mathrm{Reg}}^r(\kappa, \epsilon) = \left\{ \sup_{t\in[0,\, T]} \sup_{x\in\mathbb{R}_+} \bar{\mathcal{Z}}^r(t)([x, x+\kappa]) \leq \epsilon/5 \right\}.$$

By (71) and Lemma 5.5, for each $\epsilon > 0$ and $\eta > 0$, there exists a $\kappa > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_E^r \cap \Omega_{\mathrm{Reg}}^r(\kappa, \epsilon)) > 1 - \eta.$$

In the remainder of the proof, we set

$$\delta = \min(\epsilon/5\lambda, \kappa\epsilon/5, \epsilon^2/25)$$

and all random quantities with index r are evaluated at a fixed sample path $\omega \in \Omega_E^r \cap \Omega_{\mathrm{Reg}}^r(\kappa, \epsilon)$. For $0 \leq s \leq t \leq T$ with $t - s < \delta$, consider the following two cases.

Case 1. If $\inf_{\tau\in[s,t]} \bar{X}^r(\tau) < \epsilon/5$, let

$$t_0 = \inf\{\tau \in [s,t]\colon \bar{X}^r(\tau) \le \epsilon/5\}.$$

By right continuity, $\bar{X}^r(t_0) \le \epsilon/5$. We only need consider large enough r such that $\epsilon/5$ is smaller than K^r/r (which converges to $K > 0$ as $r \to \infty$ by condition (40)). On the interval $[s, t_0)$, $\bar{Z}^r(\cdot)$ is larger than $\epsilon/5$, implying $\bar{S}^r(s,t_0) \le |t_0 - s|/(\epsilon/5) \le |t-s|/(\epsilon/5)$. Thus, we have

$$\bar{\mathcal{Z}}^r(s)\left(A_0 + \frac{|t-s|}{\epsilon/5}\right) \le \bar{\mathcal{Z}}^r(s)(A_0 + \bar{S}^r(s,t_0)) \le \bar{\mathcal{Z}}^r(t_0)(A_0) \le \bar{X}^r(t_0) \le \epsilon/5, \tag{84}$$

where $A_0 = (0,\infty)$ and the second inequality is because of (68). Note that $\delta \le \kappa\epsilon/5$ implies that $|t-s|/(\epsilon/5) < \kappa$. Thus,

$$\bar{Z}^r(s) = \bar{\mathcal{Z}}^r(s)(A_0) \le \bar{\mathcal{Z}}^r(s)\left(A_0 + \frac{|t-s|}{\epsilon/5}\right) + \bar{\mathcal{Z}}^r(s)((0,\kappa]) \le 2\epsilon/5,$$

where the last inequality follows from (84) and the definition of $\Omega^r_{\mathrm{Reg}}(\kappa,\epsilon)$. Thus, we have

$$\bar{\mathcal{Q}}^r(s) = \mathbf{0}, \qquad \mathbf{d}[\bar{\mathcal{Z}}^r(s), \mathbf{0}] \le 2\epsilon/5.$$

On the other hand, we have

$$\bar{X}^r(t) \le \bar{X}^r(s) + \bar{E}^r(s,t) \quad \text{for all } s \le t.$$

Because we are on the event Ω^r_E and we can choose r large enough such that $\epsilon_E(r) < \epsilon/5$, we have

$$\bar{E}^r(s,t) \le \lambda\delta + \epsilon/5 \le 2\epsilon/5, \tag{85}$$

where the last inequality is because of the choice of δ. Thus, $\bar{X}^r(t) \le \bar{X}^r(t_0) + 2\epsilon/5 = 3\epsilon/5$. Again, we only need to consider large enough r such that $\epsilon/5 + 2\epsilon/5 < K^r/r$, which gives us

$$\bar{\mathcal{Q}}^r(t) = \mathbf{0}, \qquad \mathbf{d}[\bar{\mathcal{Z}}^r(t), \mathbf{0}] \le 3\epsilon/5.$$

In summary, we have that when $|t-s| \le \delta$,

$$\mathbf{d}[\bar{\mathcal{Q}}^r(s), \bar{\mathcal{Q}}^r(t)] = 0, \qquad \mathbf{d}[\bar{\mathcal{Z}}^r(s), \bar{\mathcal{Z}}^r(t)] \le 3\epsilon/5 + 2\epsilon/5 = \epsilon.$$

Case 2. If $\inf_{\tau\in[s,t]} \bar{X}^r(\tau) \ge \epsilon/5$, then $\inf_{\tau\in[s,t]} \bar{Z}^r(\tau) \ge \epsilon/5$. Therefore,

$$\bar{S}^r(s,t) \le \frac{t-s}{\epsilon/5} \le \frac{\delta}{\epsilon/5} \le \min(\kappa, \epsilon/5) \tag{86}$$

by the choice of δ. The number of jobs that enter the server during time interval $(s,t]$ is

$$\bar{B}^r(s,t) \le \bar{E}^r(s,t) + \bar{\mathcal{Z}}^r(s)([0, \bar{S}^r(s,t)]) \le 3\epsilon/5 \tag{87}$$

by (85), the choice of δ, and the definition of Ω^r_{Reg}. By the dynamic Equation (67), we have

$$|\bar{\mathcal{Q}}^r(s)(A) - \bar{\mathcal{Q}}^r(t)(A)| \le \max(\bar{E}^r(s,t), \bar{B}^r(s,t)) \le 3\epsilon/5$$

for any Borel set A. Thus,

$$\mathbf{d}[\bar{\mathcal{Q}}^r(s), \bar{\mathcal{Q}}^r(t)] \le 3\epsilon/5.$$

By the dynamic Equation (68),

$$\bar{\mathcal{Z}}^r(t)(A) \le \bar{\mathcal{Z}}^r(s)(A + \bar{S}^r(s,t)) + \bar{B}^r(s,t).$$

By (86), $A + \bar{S}^r(s,t) \subset A^{3\epsilon/5}$, where A^a is the a-enlargement of the set A as defined in §1.1. Thus, by (87),

$$\bar{\mathcal{Z}}^r(t)(A) \le \bar{\mathcal{Z}}^r(s)(A^{3\epsilon/5}) + 3\epsilon/5 \quad \text{for any Borel set } A.$$

By Property (ii) in Billingsley [3, p. 72], we have $\mathbf{d}[\bar{\mathcal{Z}}^r(s), \bar{\mathcal{Z}}^r(t)] \le 3\epsilon/5$. □

For any sequences $\{\kappa_i\}$ and $\{\delta_i\}$ of positive numbers, consider the following set

$$\left\{\sup_{t\in[0,T]}\sup_{x\in\mathbb{R}_+}\bar{\mathcal{Q}}^r(t)([x,x+\kappa_j])\le\frac{1}{j}\right\}\cap\left\{\sup_{t\in[0,T]}\sup_{x\in\mathbb{R}_+}\bar{\mathcal{Z}}^r(t)([x,x+\kappa_j])\le\frac{1}{j}\right\}$$
$$\cap\left\{\max(\mathbf{w}_T(\bar{\mathcal{Q}}^r(\cdot),\delta_j),\mathbf{w}_T(\bar{\mathcal{Z}}^r(\cdot),\delta_j))\le\frac{1}{j}\right\}.$$

Denote the two sequences $\{\kappa_j\}$ and $\{\delta_j\}$ by $\mathcal{S}$. To emphasize the dependency on $\mathcal{S}$ and j, denote the above event by $\Omega_R^r(\mathcal{S},j)$. By Lemmas 5.5, 5.6, and 5.7, for any $\eta>0$, there exists an $\mathcal{S}$ such that

$$\liminf_{r\to\infty}\mathbb{P}^r(\Omega_R^r(\mathcal{S},j))\ge 1-\frac{\eta/2}{2^j}\quad\text{for each } j\in\mathbb{N}.$$

For any finite number $n\in\mathbb{N}$, by the above inequality, we have

$$\liminf_{r\to\infty}\mathbb{P}^r\left(\bigcap_{j=1}^n\Omega_R^r(\mathcal{S},j)\right)\ge 1-\eta/2.$$

Let $r(n)$ denote the smallest number such that

$$\mathbb{P}^r\left(\bigcap_{j=1}^n\Omega_R^r(\mathcal{S},j)\right)\ge 1-\eta,\quad\text{for all } r\ge r(n).\tag{88}$$

It is clear that $r(\cdot)$ is a function defined on $\mathbb{Z}_+$ and it is nondecreasing (because $\bigcap_{j=1}^n\Omega_R^r(\mathcal{S},j)\subset\bigcap_{j=1}^{n'}\Omega_R^r(\mathcal{S},j)$ for any $n<n'$). Let

$$n(r)=\sup\{\{n\in\mathbb{Z}_+\colon r(n)\le r\}\cup\{0\}\}.$$

(From the definition, we see that $n(r)$ is allowed to be infinite, for example, when the function $r(\cdot)$ has an upper bound.) In fact, $n(\cdot)$ can be viewed as the "inverse" of $r(\cdot)$. It is clear that $n(\cdot)$ is an nondecreasing. We claim that $\lim_{r\to\infty}n(r)=\infty$. The reason is as follows: For any $n_0>0$, there exists $r_0=r(n_0)$ such that $n(r)\ge n_0$ for all $r\ge r_0$. Now, define

$$\Omega_R^r(\mathcal{S})=\bigcap_{j=1}^{n(r)}\Omega_R^r(\mathcal{S},j).$$

Note that $\Omega_R^r(\mathcal{S})$ is not empty for all large enough r (because $n(r)>1$ for all large enough r) and, in this case,

$$\mathbb{P}^r(\Omega_R^r(\mathcal{S}))\ge 1-\eta.$$

Thus, we conclude that

$$\liminf_{r\to\infty}\mathbb{P}^r(\Omega_R^r(\mathcal{S}))\ge 1-\eta.\tag{89}$$

Now, denote

$$\Omega^r(M,\mathcal{S})=\Omega_E^r\cap\Omega_B^r(M)\cap\Omega_{\mathrm{GC}}^r(M)\cap\Omega_R^r(\mathcal{S}).$$

For any r, the rth system is defined on the probability space $(\Omega^r,\mathbb{P}^r,\mathcal{F}^r)$. The stochastic processes $\mathcal{Q}^r(\cdot)$ and $\mathcal{Z}^r(\cdot)$ are actually measurable functions on Ω^r. From now on, we will explicitly write some statements down in the form of $\mathcal{Q}^r(\omega,\cdot)$ and $\mathcal{Z}^r(\omega,\cdot)$ to indicate that they are evaluated at the sample path $\omega\in\Omega^r$. We are now ready to present the precompactness result.

THEOREM 5.1. *Assume* (36)–(43). *Fix $T>0$. For all $\eta>0$, the exists a constant $M>0$ and an $\mathcal{S}$ such that*

$$\liminf_{r\to\infty}\mathbb{P}^r(\Omega^r(M,\mathcal{S}))\ge 1-\eta.\tag{90}$$

Any sequence of functions $\{(\bar{\mathcal{Q}}^{r_n}(\omega^{r_n},\cdot),\bar{\mathcal{Z}}^{r_n}(\omega^{r_n},\cdot))\}_{n\in\mathbb{N}}$ with $\omega^{r_n}\in\Omega^{r_n}(M,\mathcal{S})$ for each $n\in\mathbb{N}$ and $\{r_n\}_{n\in\mathbb{N}}$ increasing to infinity has a subsequence $\{(\bar{\mathcal{Q}}^{r_{n_i}}(\omega^{r_{n_i}},\cdot),\bar{\mathcal{Z}}^{r_{n_i}}(\omega^{r_{n_i}},\cdot))\}_{i\in\mathbb{N}}$ such that

$$\sup_{t\in[0,T]}\mathbf{d}[(\bar{\mathcal{Q}}^{r_{n_i}}(\omega^{r_{n_i}},t),\bar{\mathcal{Z}}^{r_{n_i}}(\omega^{r_{n_i}},t)),(\tilde{\mathcal{Q}}(t),\tilde{\mathcal{Z}}(t))]\to 0\quad\text{as } i\to\infty$$

for some process $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))$, which is continuous.

PROOF. The probability inequality follows immediately from (71), (76), (77), and (89).

The space $\mathbf{M}_1 \times \mathbf{M}_2$ endowed with the metric $\mathbf{d}$ (defined in §1.1) is complete. Lemma 5.4 verifies condition (a) in Theorem 3.6.3 of Ethier and Kurtz [11]. For any $\epsilon > 0$, there exists a j_0 such that $1/j < \epsilon$ for all $j \geq j_0$. Because we are on the event $\Omega_R^r(\mathcal{S})$, we have that when $\delta \leq \delta_{j_0}$ and r is large enough such that $n(r) > j_0$,

$$\max(\mathbf{w}_T(\mathcal{Q}^r(\omega^r, \cdot), \delta), \mathbf{w}_T(\mathcal{Z}^r(\omega^r, \cdot), \delta)) < \epsilon \tag{91}$$

for any $\omega^r \in \Omega^r(M, \mathcal{S})$. This verifies condition (b) in Theorem 3.6.3 of Ethier and Kurtz [11]. Thus, the sequence $\{(\bar{\mathcal{Q}}^{r_n}(\omega^{r_n}, \cdot), \bar{\mathcal{Z}}^{r_n}(\omega^{r_n}, \cdot))\}_{n\in\mathbb{N}}$ is precompact in the space $\mathbf{D}([0, T], \mathbf{M}_1 \times \mathbf{M}_2)$ endowed with the Skorohod J_1 topology. In other words, there is a convergent subsequence. The limit of this subsequence is continuous by the oscillation bound (91). Thus, convergence in the Skorohod J_1 topology is the same as convergence in the uniform metric defined in §1.1. □

6. Functional law of large number limit. Let $\mathcal{D}_T(M, \mathcal{S})$ denote the set of limits of all convergent subsequences of the sequences in Theorem 5.1. It is clear that $\mathcal{D}_T(M, \mathcal{S})$ is a nonempty subset of elements in the space $\mathbf{D}([0, T], \mathbf{M}_1 \times \mathbf{M}_2)$. We have the following result (Theorem 6.1) about the set $\mathcal{D}_T(M, \mathcal{S})$. The proof of Theorem 3.3, which builds on this result, will be provided at the end of the section.

THEOREM 6.1. *$\mathcal{D}_T(M, \mathcal{S})$ contains only one element, which is the unique fluid model solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ with initial condition (ξ^*, μ^*) restricted on the interval $[0, T]$.*

To better structure the proof, we first present three auxiliary lemmas (Lemmas 6.1, 6.2, and 6.3), which characterize any fixed element $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ in the set $\mathcal{D}_T(M, \mathcal{S})$. By the definition of $\mathcal{D}_T(M, \mathcal{S})$, for any $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathcal{D}_T(M, \mathcal{S})$, there exists a sequence $\{r_n\}$ that goes to ∞ and $\omega^{r_n} \in \Omega^{r_n}(M, \mathcal{S})$ for each r_n such that

$$\sup_{t\in[0,T]} \mathbf{d}[(\bar{\mathcal{Q}}^{r_n}(\omega^{r_n}, t), \bar{\mathcal{Z}}^{r_n}(\omega^{r_n}, t)), (\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t))] \to 0 \quad \text{as } n \to \infty.$$

With a slight abuse of notation, we drop the parameter ω^{r_n} for simplicity in the proofs of all the following three lemmas. We then have

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \mathbf{d}[\bar{\mathcal{Q}}^{r_n}(t), \tilde{\mathcal{Q}}(t)] = 0, \tag{92}$$

$$\lim_{n\to\infty} \sup_{t\in[0,T]} \mathbf{d}[\bar{\mathcal{Z}}^{r_n}(t), \tilde{\mathcal{Z}}(t)] = 0. \tag{93}$$

LEMMA 6.1. *Assume (36)–(43). For any point $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathcal{D}_T(M, \mathcal{S})$, both $\tilde{\mathcal{Q}}(t)$ and $\tilde{\mathcal{Z}}(t)$ are atom-free for all $t \in [0, T]$.*

PROOF. For any $y \geq 0$ and $\kappa_1 > 0$, because $[y - \kappa_1, y + 2\kappa_1]$ is the κ_1-enlargement (cf. §1.1) of the set $[y, y + \kappa_1]$, by (92) and the definition of Prohorov metric, we have

$$\tilde{\mathcal{Q}}(t)([y, y + \kappa_1]) \leq \bar{\mathcal{Q}}^{r_n}(t)([y - \kappa_1, y + 2\kappa_1]) + \kappa_1$$

for all large n. Because we are on the event $\Omega^{r_n}(M, \mathcal{S})$, in particular $\Omega_R^{r_n}(\mathcal{S})$, for any $\epsilon > 0$, we can choose κ_1 small enough such that

$$\bar{\mathcal{Q}}^{r_n}(t)([y - \kappa_1, y + 2\kappa_1]) < \epsilon/2$$

for all large n. When making κ_1 small, we can also choose $\kappa_1 < \epsilon/2$. This gives that

$$\tilde{\mathcal{Q}}(t)([y, y + \kappa_1]) < \epsilon.$$

This proves that $\tilde{\mathcal{Q}}(t)$ is atom-free for any $t \in [0, T]$. The proof for $\tilde{\mathcal{Z}}(t)$ follows in exactly the same way. □

LEMMA 6.2. *Assume (36)–(43). Fix any point $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathcal{D}_T(M, \mathcal{S})$ and constants $a, b \in [0, T]$ with $a < b$. If*

$$\inf_{t\in[a,b]} \tilde{Z}(t) > 0, \tag{94}$$

then $(\tilde{\mathcal{Q}}(a), \tilde{\mathcal{Z}}(a)) \in \mathcal{I}$ and $(\tilde{\mathcal{Q}}(a + \cdot), \tilde{\mathcal{Z}}(a + \cdot))$ is the solution to the fluid model (K, λ, ν) with initial condition $(\tilde{\mathcal{Q}}(a), \tilde{\mathcal{Z}}(a))$ on the interval $[0, b - a]$.

PROOF. Define $\tilde{Q}(\cdot)$, $\tilde{Z}(\cdot)$, $\tilde{B}(\cdot)$, and $\tilde{S}(\cdot,\cdot)$ in the same way as (13)–(20). Then, (92) and (93) imply that

$$\lim_{n\to\infty}\sup_{t\in[0,T]}|\bar{Q}^{r_n}(t)-\tilde{Q}(t)|=0, \tag{95}$$

$$\lim_{n\to\infty}\sup_{t\in[0,T]}|\bar{Z}^{r_n}(t)-\tilde{Z}(t)|=0, \tag{96}$$

$$\lim_{n\to\infty}\sup_{t\in[0,T]}|\bar{B}^{r_n}(t)-\tilde{B}(t)|=0. \tag{97}$$

By (94) and (96),

$$\sup_{a\le t\le b}\left|\frac{1}{\bar{Z}^{r_n}(t)}-\frac{1}{\tilde{Z}(t)}\right|\to 0 \quad \text{as } n\to\infty.$$

Thus, for each $\epsilon>0$, there exists an $n_0>0$ such that

$$\sup_{a\le s<t\le b}|\bar{S}^{r_n}(s,t)-\tilde{S}(s,t)|<\epsilon, \quad \text{for all } n>n_0. \tag{98}$$

Because for all r_n, $(\bar{\mathcal{Q}}^{r_n}(\cdot),\bar{\mathcal{Z}}^{r_n}(\cdot))$ satisfies the LPS policy constraints (9) and (10) and $K^{r_n}/r_n\to K$ as $n\to\infty$, the limit $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))$ also satisfies (24) and (25). It is then clear that $(\tilde{\mathcal{Q}}(a),\tilde{\mathcal{Z}}(a))$ is a valid initial condition. By the same argument, $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))$ also satisfies (23). Now, it only remains to show that $(\tilde{\mathcal{Q}}(a+\cdot),\tilde{\mathcal{Z}}(a+\cdot))$ satisfies the fluid dynamic Equations (21) and (22) on the interval $[0,b-a]$.

By (67), for any Borel set $A\subset\mathbb{R}_+$ and $t\ge 0$,

$$\bar{\mathcal{Q}}^{r_n}(a+t)(A)=\bar{\mathcal{Q}}^{r_n}(a)(A)+I_0^{r_n}(A)-I_1^{r_n}(A), \tag{99}$$

where

$$I_0^{r_n}(A)=\frac{1}{r_n}\sum_{i=r_n\bar{E}^{r_n}(a)+1}^{r_n\bar{E}^{r_n}(a+t)}\delta_{v_i^{r_n}}(A),$$

$$I_1^{r_n}(A)=\frac{1}{r_n}\sum_{i=r_n\bar{B}^{r_n}(a)+1}^{r_n\bar{B}^{r_n}(a+t)}\delta_{v_i^{r_n}}(A).$$

To verify (21), consider the following difference for any $y\ge 0$ (recall that $A_y=(y,\infty)$),

$$\begin{aligned}
&|\tilde{\mathcal{Q}}(a+t)(A_y)-(\tilde{\mathcal{Q}}(a)(A_y)+[\tilde{Q}(a+t)-\tilde{Q}(a)]\nu(A_y))|\\
&\quad\le|\tilde{\mathcal{Q}}(a+t)(A_y)-\bar{\mathcal{Q}}^{r_n}(a+t)(A_y)|\\
&\qquad+|\bar{\mathcal{Q}}^{r_n}(a+t)(A_y)-(\bar{\mathcal{Q}}^{r_n}(a)(A_y)+I_0^{r_n}(A_y)-I_1^{r_n}(A_y))|\\
&\qquad+|(\bar{\mathcal{Q}}^{r_n}(a)(A_y)+I_0^{r_n}(A_y)-I_1^{r_n}(A_y))-(\tilde{\mathcal{Q}}(a)(A_y)+[\tilde{Q}(a+t)-\tilde{Q}(a)]\nu(A_y))|\\
&\quad\le|\tilde{\mathcal{Q}}(a+t)(A_y)-\bar{\mathcal{Q}}^{r_n}(a+t)(A_y)|+|\tilde{\mathcal{Q}}(a)(A_y)-\bar{\mathcal{Q}}^{r_n}(a)(A_y)|\\
&\qquad+|[\tilde{Q}(a+t)\tilde{Q}(a)]\nu(A_y)-I_0^{r_n}(A_y)+I_1^{r_n}(A_y)|,
\end{aligned} \tag{100}$$

where the first inequality is because of triangle inequality and the second one is because of (99) and another application of triangle inequality. According to Lemma 6.1, the set A_y is a $\tilde{\mathcal{Q}}(a+t)$-continuity set (i.e., a set whose boundary has zero mass under the measure). By Property (iii) of Billingsley [3, p. 72], the convergence of $\bar{\mathcal{Q}}^{r_n}(a+t)$ to $\tilde{\mathcal{Q}}(a+t)$ in the Prohorov metric implies weak convergence. By Portmanteau theorem (cf. Theorem 2.1 in Billingsley [3]), weak convergence implies $\bar{\mathcal{Q}}^{r_n}(a+t)(A)\to\tilde{\mathcal{Q}}(a+t)(A)$ for all $\tilde{\mathcal{Q}}(a+t)$-continuity set A. This implies that each of the first two terms on the right-hand side of (99) can be bounded by ϵ for all large n. Now, let us study the third term. Let $\tilde{E}(\cdot)=\tilde{B}(\cdot)+\tilde{Q}(\cdot)$, so $\tilde{E}(\cdot)$ is the limit of $\bar{E}^r(\cdot)$. (In fact, $\tilde{E}(\cdot)=\lambda\cdot$ as proved in §5.1. However, it is not needed here.) By triangle inequality, we have that

$$\begin{aligned}
&|[\tilde{Q}(a+t)-\tilde{Q}(a)]\nu(A_y)-I_0^{r_n}(A_y)+I_1^{r_n}(A_y)|\\
&\quad=|[\tilde{E}(a+t)-\tilde{E}(a)]\nu(A_y)-I_0^{r_n}(A_y)-[\tilde{B}(a+t)-\tilde{B}(a)]\nu(A_y)+I_1^{r_n}(A_y)|\\
&\quad\le|[\tilde{E}(a+t)-\tilde{E}(a)]\nu(A_y)-I_0^{r_n}(A_y)|+|\tilde{B}(a+t)-\tilde{B}(a)]\nu(A_y)-I_1^{r_n}(A_y)|.
\end{aligned}$$

Note that

$$\begin{aligned}&|[\tilde{E}(a+t)-\tilde{E}(a)]\nu(A_y)-I_0^{r_n}(A_y)|\\&\quad\le[\tilde{E}(a+t)-\tilde{E}(a)]|\nu(A_y)-\nu^{r_n}(A_y)|+|[\tilde{E}(a+t)-\tilde{E}(a)]\nu^{r_n}(A_y)-I_0^{r_n}(A_y)|.\end{aligned}$$

Again, because ν is atom-free (by condition (39)), A_y is a ν-continuity set. Thus, $|\nu(A_y)-\nu^{r_n}(A_y)|\le\epsilon$ for all large n. Because we restrict our sample path to be in the event $\Omega^{r_n}(M,\mathcal{S})$ and hence in $\Omega_E^{r_n}\cap\Omega_B^{r_n}(M)$ for each n, the limits $\tilde{E}(\cdot)$ and $\tilde{B}(\cdot)$ have an upper bound $M+2\lambda T$ and a lower bound $-M$ on the interval $[0,T]$. Thus, the first term in the above can be bounded by $(M+2\lambda T)\epsilon$ for all large n. Note that

$$\begin{aligned}&|[\tilde{E}(a+t)-\tilde{E}(a)]\nu^{r_n}(A_y)-I_0^{r_n}(A_y)|\\&\quad\le|\bar{E}^{r_n}(a+t)-\tilde{E}(a+t)|+|\bar{E}^{r_n}(a)-\tilde{E}(a)|\\&\qquad+\left|\frac{1}{r_n}\sum_{i=r_n\tilde{E}(a)+1}^{r_n\tilde{E}(a+t)}\delta_{v_i^{r_n}}(A_y)-[\tilde{E}(a+t)-\tilde{E}(a)]\nu^{r_n}(A_y)\right|.\end{aligned}$$

Because $\tilde{E}(\cdot)$ is the limit of $\bar{E}^{r_n}(\cdot)$, each of the first two terms is bounded by ϵ for all large n. Because we restrict our sample path to be in the event $\Omega^{r_n}(M,\mathcal{S})$ and hence in $\Omega_{\mathrm{GC}}^{r_n}(M)$ for all n, the last term in the above can be bounded above by ϵ for all large n. Thus, we conclude that

$$|[\tilde{E}(a+t)-\tilde{E}(a)]\nu(A_y)-I_0^{r_n}(A_y)|\le(M+2\lambda T+3)\epsilon$$

for all large n. Using exactly the same argument, we can show that

$$|[\tilde{B}(a+t)-\tilde{B}(a)]\nu(A_y)-I_1^{r_n}(A_y)|\le(M+2\lambda T+3)\epsilon$$

for all large n. In summary, the right side of (100) is bounded by $(2M+4\lambda T+8)\epsilon$ for all large n. Because $\epsilon>0$ is arbitrary, the left side of (100) must be 0 and, thus, the fluid dynamic Equation (21) is verified.

By (68), for all Borel set $A\subset\mathbb{R}_+$ and $t\ge0$,

$$\bar{\mathcal{Z}}^{r_n}(a+t)(A)=\bar{\mathcal{Z}}^{r_n}(a)(A+\bar{S}^{r_n}(a,a+t))+I_2^{r_n}(A),\tag{101}$$

where

$$I_2^{r_n}(A)=\frac{1}{r_n}\sum_{i=r_n\bar{B}^{r_n}(a)+1}^{r_n\bar{B}^{r_n}(a+t)}\delta_{v_i^r}(A+\bar{S}^{r_n}(\tau_i^{r_n},a+t)).$$

To verify (22), consider the difference

$$\begin{aligned}&\left|(\tilde{\mathcal{Z}}(a+t)(A_y)-\tilde{\mathcal{Z}}(a)(A_y+\tilde{S}(a,a+t)))-\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))\,d\tilde{B}(\tau)\right|\\&\quad\le|(\tilde{\mathcal{Z}}(a+t)(A_y)-\tilde{\mathcal{Z}}(a)(A_y+\tilde{S}(a,a+t)))\\&\qquad-(\bar{\mathcal{Z}}^{r_n}(a+t)(A_y)-\bar{\mathcal{Z}}^{r_n}(a)(A_y+\bar{S}^{r_n}(a,a+t)))|\\&\qquad+|(\bar{\mathcal{Z}}^{r_n}(a+t)(A_y)-\bar{\mathcal{Z}}^{r_n}(a)(A_y+\bar{S}^{r_n}(a,a+t)))-I_2^{r_n}(A_y)|\\&\qquad+\left|\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))\,d\tilde{B}(\tau)-I_2^{r_n}(A_y)\right|\\&\quad\le|\tilde{\mathcal{Z}}(a+t)(A_y)-\bar{\mathcal{Z}}^{r_n}(a+t)(A_y)|\\&\qquad+|\tilde{\mathcal{Z}}(a)(A_y+\tilde{S}(a,a+t))-\bar{\mathcal{Z}}^{r_n}(a)(A_y+\bar{S}^{r_n}(a,a+t))|\\&\qquad+\left|\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))\,d\tilde{B}(\tau)-I_2^{r_n}(A_y)\right|,\end{aligned}\tag{102}$$

where the first inequality is because of triangle inequality, and the second one is because of (101) and another application of triangle inequality. By Lemma 6.1, the measure $\tilde{\mathcal{Z}}(t+a)$ is also atom-free. Following the same

argument as the one for $\tilde{\mathcal{Q}}(a)$, the first term on the right-hand side in (102) is bounded by ϵ for all large n. For any $y \geq 0$ and $\kappa > 0$,

$$\begin{aligned}
&\tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty)) - \bar{\mathcal{Z}}^{r_n}(a)((y+\bar{S}^{r_n}(a,a+t),\infty)) \\
&\quad \leq \tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty)) - \bar{\mathcal{Z}}^{r_n}(a)((y+\tilde{S}(a,a+t)+\kappa,\infty)) \\
&\quad \leq \tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty)) - \bar{\mathcal{Z}}^{r_n}(a)((y+\tilde{S}(a,a+t)-\kappa,\infty)) \\
&\qquad + \bar{\mathcal{Z}}^{r_n}(a)([y+\tilde{S}(a,a+t)-\kappa, y+\tilde{S}(a,a+t)+\kappa]) \\
&\quad \leq \kappa + \bar{\mathcal{Z}}^{r_n}(a)([y+\tilde{S}(a,a+t)-\kappa, y+\tilde{S}(a,a+t)+\kappa])
\end{aligned}$$

for all large n, where the first inequality is because of (98), the second inequality is because of algebra, and the last inequality is because of (93) and the definition of Prohorov metric. Because we restrict our sample path to be in the event $\Omega^{r_n}(M,\mathcal{S})$ and hence in $\Omega_R^{r_n}(\mathcal{S})$ for all n, we can choose κ small enough (less than ϵ) to make the second term on the right-hand side of the above less than ϵ. Thus, we have

$$\tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty)) - \bar{\mathcal{Z}}^{r_n}(a)((y+\bar{S}^{r_n}(a,a+t),\infty)) \leq 2\epsilon.$$

On the other side, for any $y \geq 0$ and $\kappa > 0$,

$$\begin{aligned}
&\bar{\mathcal{Z}}^{r_n}(a)((y+\bar{S}^{r_n}(a,a+t),\infty)) - \tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty)) \\
&\quad \leq \bar{\mathcal{Z}}^{r_n}(a)((y+\tilde{S}(a,a+t)-\kappa,\infty)) - \tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty)) \\
&\quad \leq \bar{\mathcal{Z}}^{r_n}(a)([y+\tilde{S}(a,a+t)-\kappa, y+\tilde{S}(a,a+t)+\kappa]) \\
&\qquad + \bar{\mathcal{Z}}^{r_n}(a)((y+\tilde{S}(a,a+t)+\kappa,\infty)) - \tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty))
\end{aligned}$$

for all large n, where the first inequality is because of (98) and the second inequality is because of algebra. By the same argument, we can show that

$$\bar{\mathcal{Z}}^{r_n}(a)((y+\bar{S}^{r_n}(a,a+t),\infty)) - \tilde{\mathcal{Z}}(a)((y+\tilde{S}(a,a+t),\infty)) \leq 2\epsilon.$$

This implies that the second term on the right-hand side of (102) is bounded by 2ϵ. To control the third term, define

$$I_2^{r_n}(A) = \sum_{j=0}^{N-1} I_{2,j}^{r_n}(A),$$

where $0 = t_0 < \cdots < t_{N-1} = t$ is a partition of the interval $[0,t]$ with $\delta = \max_j |t_{j+1} - t_j|$ and

$$I_{2,j}^{r_n}(A) = \frac{1}{r_n} \sum_{i=r_n\bar{B}^{r_n}(a+t_j)+1}^{r_n\bar{B}^{r_n}(a+t_{j+1})} \delta_{v_i^r}(A + \bar{S}^{r_n}(\tau_i^{r_n}, a+t)).$$

Recall that $\tau_i^{r_n}$ is the time that the ith job starts service in the r_nth system, so on each subinterval $[a+t_j, a+t_{j+1}]$ those i's to be summed must satisfy $a+t_j \leq \tau_i^{r_n} \leq a+t_{j+1}$. This implies that

$$\bar{S}^{r_n}(a+t_{j+1}, a+t) \leq \bar{S}^{r_n}(\tau_i^{r_n}, a+t) \leq \bar{S}^{r_n}(a+t_j, a+t).$$

By the uniform convergence (98), we have for all large n,

$$y - \epsilon + \tilde{S}(a+t_{j+1}, a+t) \leq y + \bar{S}^{r_n}(\tau_i^{r_n}, a+t) \leq y + \epsilon + \tilde{S}(a+t_j, a+t).$$

Because we are on the event $\Omega^{r_n}(M,\mathcal{S})$ (which is defined at the end of §5), for $\epsilon > 0$, there exists an n_1 such that for all $n > n_1$ and $j = 0, \ldots, N-1$,

$$\begin{aligned}
I_{2,j}^{r_n}(A_y) &\geq \bar{B}^{r_n}(a+t_j, a+t_{j+1})\nu^{r_n}(A_y + \epsilon + \tilde{S}(a+t_j, a+t)) - \epsilon, \\
&\geq \tilde{B}(a+t_j, a+t_{j+1})\nu^{r_n}(A_y + \epsilon + \tilde{S}(a+t_j, a+t)) - 2\epsilon, \\
&\geq \tilde{B}(a+t_j, a+t_{j+1})\nu(A_y + \tilde{S}(a+t_j, a+t)) - (2M + 2\lambda T + 2)\epsilon,
\end{aligned}$$

where the above three inequalities are because of the fact that we are on the event $\Omega^r_{\mathrm{GC}}(M)$, (97) and to the definition of Prohorov metric, respectively. Note that the above $2M+2\lambda T$ comes from $\tilde{B}(s,t)=\tilde{B}(t)-\tilde{B}(s)<2M+2\lambda T$ because $\tilde{B}(\cdot)$ has lower bounded $-M$ and upper bound $(M+2\lambda T)$ (again, because we are on the event $\Omega^{r_n}_B(M)$). By the same reason, for all $n>n_1$ and $j=0,\ldots,N-1$,

$$I^{r_n}_{2,j}(A_y)\le \tilde{B}(a+t_j,a+t_{j+1})\nu(A_y+\tilde{S}(a+t_{j+1},a+t))+(2M+2\lambda T+2)\epsilon.$$

Denoting

$$I_{L,\delta}(A_y)=\sum_{j=0}^{N-1}\tilde{B}(a+t_j,a+t_{j+1})\nu(A_y+\tilde{S}(a+t_j,a+t)),$$

$$I_{U,\delta}(A_y)=\sum_{j=0}^{N-1}\tilde{B}(a+t_j,a+t_{j+1})\nu(A_y+\tilde{S}(a+t_{j+1},a+t)),$$

we have that

$$I_{L,\delta}(A_y)-N(2M+2\lambda T+2)\epsilon\le I^{r_n}_2(A_y)\le I_{U,\delta}(A_y)+N(2M+2\lambda T+2)\epsilon. \tag{103}$$

It is clear that $I_{L,\delta}(A_y)$ and $I_{U,\delta}(A_y)$ are the Riemann lower sum and upper sum of the integration $\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))\,d\tilde{B}(\tau)$, respectively. This means that for all δ small enough,

$$I_{L,\delta}(A_y)\le\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))\,d\tilde{B}(\tau)\le I_{U,\delta}(A_y). \tag{104}$$

It then follows from (103) and (104) that

$$\left|\int_a^{a+t}\nu(A_y+\tilde{S}(\tau,a+t))\,d\tilde{B}(\tau)-I^{r_n}_2(A_y)\right|\le[I_{U,\delta}(A_y)-I_{L,\delta}(A_y)]+2N(2M+2\lambda T+2)\epsilon. \tag{105}$$

For any $\epsilon_1>0$, we first choose δ small enough (therefore, N is chosen) to make the first term in the upper bound of (105) less than $\epsilon_1/2$, and then choose ϵ small enough to make the second term in the upper bound of (105) less than $\epsilon_1/2$. Thus, the third term on the right-hand side of (102) is bounded by ϵ_1. In summary, the right-hand side of (102) is bounded by $3\epsilon+\epsilon_1$ for all large n. Because $\epsilon,\epsilon_1>0$ can be arbitrarily small, the left-hand side of (102) must be zero. Thus, (22) is satisfied. □

Lemma 6.3. *Assume* (36)–(43). *Fix a point* $(\tilde{\mathcal{Q}}(\cdot),\tilde{\mathcal{Z}}(\cdot))\in\mathcal{D}_T(M,\mathcal{S})$ *and a constant* $t_0\in[0,T]$. *If* $(\tilde{\mathcal{Q}}(t_0),\tilde{\mathcal{Z}}(t_0))=(\mathbf{0},\mathbf{0})$, *then*

$$(\tilde{\mathcal{Q}}(t),\tilde{\mathcal{Z}}(t))=(\mathbf{0},\mathbf{0}),\quad \textit{for all } t\in[t_0,T] \tag{106}$$

when $\rho\le 1$;

$$\inf_{t\in[t_1,T]}\tilde{Z}(t)>0,\quad \textit{for all } t_1\in(t_0,T] \tag{107}$$

when $\rho>1$. *If* $(\tilde{\mathcal{Q}}(t_0),\tilde{\mathcal{Z}}(t_0))\ne(\mathbf{0},\mathbf{0})$ *and* $\rho>1$, *then*

$$\inf_{t\in[t_0,T]}\tilde{Z}(t)>0. \tag{108}$$

Proof. The assumption $(\tilde{\mathcal{Q}}(t_0),\tilde{\mathcal{Z}}(t_0))=(\mathbf{0},\mathbf{0})$ implies that

$$\bar{\mathcal{Z}}^{r_n}(t_0)\to\mathbf{0}\quad\text{as } n\to\infty. \tag{109}$$

Note that for any constant $a>0$, the workload at time t_0 satisfies

$$\begin{aligned}\bar{W}^{r_n}(t_0)&=\langle\chi,\bar{\mathcal{Z}}^{r_n}(t_0)\rangle\\&=\langle\chi 1_{[0,a]},\bar{\mathcal{Z}}^{r_n}(t_0)\rangle+\langle\chi 1_{[a,\infty)},\bar{\mathcal{Z}}^{r_n}(t_0)\rangle\\&\le a\langle 1,\bar{\mathcal{Z}}^{r_n}(t_0)\rangle+\frac{1}{a^p}\langle\chi^{1+p},\bar{\mathcal{Z}}^{r_n}(t_0)\rangle.\end{aligned}$$

By the definition of $\Omega^{r_n}_B(M)$, $\langle\chi^{1+p},\bar{\mathcal{Z}}^{r_n}(t_0)\rangle<M$. For any $\epsilon>0$, we first choose a large enough such that $M/a^p<\epsilon/2$. By (109), we can then choose n large enough such that $a\langle 1,\bar{\mathcal{Z}}^{r_n}(t_0)\rangle\le\epsilon/2$. This implies that

$$\bar{W}^{r_n}(t_0)=\langle\chi,\bar{\mathcal{Z}}^{r_n}(t_0)\rangle\to 0\quad\text{as } n\to\infty.$$

By Proposition 3.2,

$$\bar{W}^{r_n}(\cdot) \to \bar{W}(\cdot),$$

where $\bar{W}(t) = (w^* - (1-\rho)t)^+$ for $t \geq 0$. This means that $\bar{W}(t_0) = 0$.

If $\rho \leq 1$, then $\bar{W}(t) = 0$ for all $t \geq t_0$. This means that for each $t \geq t_0$, $\bar{W}^{r_n}(t) \to 0$ as $n \to \infty$. For any $\kappa > 0$, we have the following inequality:

$$\bar{Z}^{r_n}(t) \leq \bar{\mathcal{Z}}^{r_n}(t)(0, \kappa] + \frac{\bar{W}^{r_n}(t)}{\kappa}.$$

Because we are on the event $\Omega_R^{r_n}(\mathcal{S})$ (which is defined at the end of §5), we can choose κ small enough such that

$$\bar{Z}^{r_n}(t) \leq \epsilon + \frac{\bar{W}^{r_n}(t)}{\kappa},$$

where the second term on the right-hand side in the above can be made smaller than ϵ by taking n large enough. This implies that $\tilde{Z}(t) = 0$, which means $(\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$.

If $\rho > 1$, then for any $t \in [t_1, T]$, we have

$$\bar{W}(t) \geq (\rho - 1)(t_1 - t_0) \stackrel{\Delta}{=} \alpha_1.$$

Because on the event $\Omega_B^r(M)$ (which is defined in §5.1), $\langle \chi^{1+p}, \bar{\mathcal{Z}}^{r_n}(t) \rangle < M$ for all $t \in [0, T]$, for any $\epsilon > 0$, there exists a $c_0 > 0$ such that

$$\langle \chi 1_{(c_0, \infty)}, \bar{\mathcal{Z}}^{r_n}(t) \rangle < \epsilon \quad \text{for all } t \in [0, T] \text{ and } n \geq 0.$$

This implies that for all $t \in [0, T]$,

$$\begin{aligned} \bar{W}^{r_n}(t) &= \langle \chi 1_{[0, c_0]}, \bar{\mathcal{Z}}^{r_n}(t) \rangle + \langle \chi 1_{(c_0, \infty)}, \bar{\mathcal{Z}}^{r_n}(t) \rangle \\ &\leq c_0 \bar{Z}^{r_n}(t) + \epsilon. \end{aligned} \tag{110}$$

Taking $\epsilon = \alpha_1/2$, we have that $\bar{Z}^{r_n}(t) \geq \alpha_1/(2c_0)$ for all $t \in [t_1, T]$. Letting $n \to \infty$, $\tilde{Z}(t) \geq \alpha_1/(2c_0)$ for all $t \in [t_1, T]$.

The assumption $(\tilde{\mathcal{Q}}(t_0), \tilde{\mathcal{Z}}(t_0)) \neq (\mathbf{0}, \mathbf{0})$ implies that $\tilde{Z}(t_0) > 0$. If $\rho > 1$, then for any $t \in [t_0, T]$, we have

$$\bar{W}(t) = \bar{W}(t_0) + (\rho - 1)(t - t_0) \geq \bar{W}(t_0) \stackrel{\Delta}{=} \alpha_0 > 0.$$

Note that (110) holds on the interval $[0, T]$, we apply it to the interval $[t_0, T]$. Taking $\epsilon = \alpha_0/2$, we have that $\bar{Z}^{r_n}(t) \geq \alpha_0/(2c_0)$ for all $t \in [t_0, T]$. Letting $n \to \infty$, $\tilde{Z}(t) \geq \alpha_1/(2c_0)$ for all $t \in [t_0, T]$. □

Proof of Theorem 6.1. *Case* 1, $\rho > 1$. By Lemmas 6.2 and 6.3 for any $0 < t_1 \leq t$, we have that

$$\tilde{\mathcal{Q}}(t)(A_y) = \tilde{Q}(t_1)\nu(A_y) + [\tilde{Q}(t) - \tilde{Q}(t_1)]\nu(A_y), \tag{111}$$

$$\tilde{\mathcal{Z}}(t)(A_y) = \tilde{\mathcal{Z}}(t_1)(A_y + \tilde{S}(t_1, t)) + \int_{t_1}^{t} \nu(A_y + \tilde{S}(s, t))\, d\bar{B}(s) \tag{112}$$

for all $y \geq 0$. Because $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is continuous, we have that

$$(\tilde{\mathcal{Q}}(t_1), \tilde{\mathcal{Z}}(t_1)) \to (\xi^*, \mu^*) \quad \text{as } t_1 \to 0. \tag{113}$$

Thus, letting $t_1 \to 0$, (111) becomes

$$\tilde{\mathcal{Q}}(t)(A_y) = \xi^*(A_y) + [\tilde{Q}(t) - \tilde{Q}(0)]\nu(A_y).$$

Note that

$$\int_0^{t_1} \nu(A_y + \tilde{S}(s, t))\, d\bar{B}(s) \leq [\tilde{B}(t_1) - \tilde{B}(0)] = \lambda t_1 - (\tilde{Q}(t_1) - \tilde{Q}(0)),$$

which converges to 0 as $t_1 \to 0$. If $\mu^* \neq \mathbf{0}$, then $\tilde{Z}(0) > 0$. Lemma 6.3 implies that $\inf_{s \in [0, t]} \tilde{Z}(s) > 0$. This implies that

$$\tilde{S}(t_1, t) \to \tilde{S}(t) \quad \text{as } t_1 \to 0. \tag{114}$$

By (113), $\widetilde{\mathcal{Z}}(t_1) \to \mu^*$ (in the Prohorov metric) as $t_1 \to 0$. It follows from (114) that

$$\widetilde{\mathcal{Z}}(t_1)(A_y + \tilde{S}(t_1, t)) \to \mu^*(A_y + \tilde{S}(t)) \quad \text{as } t_1 \to 0.$$

If $\mu^* = \mathbf{0}$,

$$\widetilde{\mathcal{Z}}(t_1)(A_y + \tilde{S}(t_1, t)) \to 0 \quad \text{as } t_1 \to 0.$$

In both cases, letting $t_1 \to 0$, (112) becomes

$$\widetilde{\mathcal{Z}}(t)(A_y) = \mu^*(A_y + \tilde{S}(t)) + \int_0^t \nu(A_y + \tilde{S}(s, t))\, d\bar{B}(s).$$

We conclude that $(\widetilde{\mathcal{Q}}(\cdot), \widetilde{\mathcal{Z}}(\cdot))$ is the fluid model solution with initial condition (ξ^*, μ^*).

Case 2, $\rho \le 1$. Let $t_0 = \inf\{t \ge 0\colon \widetilde{\mathcal{Z}}(t) = \mathbf{0}\}$. By Lemma 6.2, for all $t \in [0, t_0)$, $(\widetilde{\mathcal{Q}}(t), \widetilde{\mathcal{Z}}(t))$ satisfies the fluid dynamic Equations (21) and (22) with initial condition (ξ^*, μ^*). By the continuity of $(\widetilde{\mathcal{Q}}(\cdot), \widetilde{\mathcal{Z}}(\cdot))$ and Lemma 6.3, $(\widetilde{\mathcal{Q}}(t), \widetilde{\mathcal{Z}}(t)) = (\mathbf{0}, \mathbf{0})$ for all $t \in [t_0, T]$. Thus, $(\widetilde{\mathcal{Q}}(\cdot), \widetilde{\mathcal{Z}}(\cdot))$ is the fluid model solution with initial condition (ξ^*, μ^*). □

Proof of Theorem 3.3. It is enough to show that for any $\eta, \epsilon > 0$,

$$\liminf_{r\to\infty} \mathbb{P}^r(\varrho[(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot)), (\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))] < \epsilon) \ge 1 - \eta,$$

where ϱ is the Skorohod metric defined in §1.1. Fix a constant $T > 0$ such that $\int_T^\infty e^{-t}\, dt < \epsilon/2$. By Theorems 5.1 and 6.1, we have that

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\varrho_T[(\bar{\mathcal{Q}}^r(\cdot), \bar{\mathcal{Z}}^r(\cdot)), (\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))] < \frac{\epsilon}{2(1 - e^{-T})}\right) \ge 1 - \eta.$$

The result follows immediately from (3). □

Appendix A. A convolution equation.

Lemma A.1. *Suppose $F(0) < 1$, $\rho > 0$ and $h(\cdot)$ is a càdlàg function. There exists a $b > 0$ (only depending on ρ and F) such that the two-side convolution Equation (50) has a unique solution $x(\cdot)$ on $[0, b]$, which is càdlàg.*

Proof. The space $\mathbf{D}([0, b], \mathbb{R})$ (all real-valued *càdlàg* functions on $[0, b]$, cf. §1.1) is a subset of the Banach space of bounded, measurable functions on $[0, b]$, equipped with the sup norm. One can check that this subset is closed in the Banach space. Thus, the space $\mathbf{D}([0, b], \mathbb{R})$ itself, equipped with the uniform metric υ_b (defined in §1.1), is complete.

Because $F(0) < 1$, there exists $b > 0$ such that

$$\kappa := \rho F_e(b) + F(b) < 1.$$

For any $y \in \mathbf{D}([0, b], \mathbb{R})$, define $\Psi(y)$ by

$$\Psi(y)(u) = h(u) + \rho \int_0^u (y(u - v) \wedge K)\, dF_e(v) + \int_0^u (y(u - v) - K)^+\, dF(v)$$

for any $u \in [0, b]$. By convention, the integration $\int_0^u y(u - v)\, dF(v)$ is interpreted to be $\int_{(0,u]} y(u - v)\, dF(v)$ (cf. Chung [5, p. 43]).

First, we show that Ψ is a mapping from $\mathbf{D}([0, b], \mathbb{R})$ to $\mathbf{D}([0, b], \mathbb{R})$. Because the function h is a *càdlàg* function, essentially we only need to show that the convolution $z(u) = \int_0^u y(u - v)\, dF(v)$ is a *càdlàg* function for any *càdlàg* function y and distribution function F. By Theorem 12.2.2 in Whitt [30], there exists a sequence of piecewise constant *càdlàg* functions y_n such that $\upsilon_b[y_n, y] \to 0$ as $n \to \infty$. By piecewise constant *càdlàg*, we mean a function of the form

$$\sum_{j=0}^{J-1} c_j 1_{[a_j, b_j)} + c_J 1_{[a_J, b]},$$

where $c_j \in \mathbb{R}$, $a_j, b_j \in [0, b]$ with $a_j < b_j$ for all $j = 0, \ldots, J - 1$ and $a_J < b$. Note that the convolution of indicator function $1_{[a_j, b_j)}$,

$$\int_0^u 1_{[a_j, b_j)}(u - v)\, dF(v),$$

equals 0 if $u < a_j$, equals $F(u-a_j)-F(0)$ if $u \in [a_j, b_j)$, and equals $F(u-a_j)-F(u-b_j)$ if $u \geq b_j$. Because F is *càdlàg*, the convolution of $1_{[a_i, b_j)}$ is also *càdlàg*. Similarly, the convolution of indicator function $1_{[a_J, b]}$,

$$\int_0^u 1_{[a_J, b]}(u-v)\, dF(v),$$

equals 0 if $u < a_J$ and equals $F(u-a_J)-F(0)$ if $u \in [a_J, b]$. Again, this convolution is a *càdlàg* function. It is now easy to see that $z_n(u) = \int_0^u y_n(u-v)\, dF(v)$ is a *càdlàg* function for each n because it is a linear combination of *càdlàg* functions. For any n, we have that

$$\begin{aligned} \upsilon_b[z_n, z] &\leq \sup_{u\in[0,\, b]} \int_0^u |y_n(u-v) - y(u-v)|\, dF(v) \\ &\leq \int_0^u \upsilon_b[y_n, y]\, dF(v) \leq F(u)\upsilon_b[y_n, y]. \end{aligned}$$

Thus, $\upsilon_b[y_n, y] \to 0$ implies that $\upsilon_b[z_n, z] \to 0$. Because the space $\mathbf{D}([0, b], \mathbb{R})$ is complete under the uniform metric, the limit z is a *càdlàg* function.

Next, we show that the mapping Ψ is a contraction. For any $y, y' \in \mathbf{D}([0, b], \mathbb{R})$, we have that

$$\begin{aligned} \upsilon_b[\Psi(y), \Psi(y')] \leq{}& \sup_{u\in[0,\, b]} \rho \int_0^u |(y(u-v)\wedge K) - (y'(u-v)\wedge K)|\, dF_e(v) \\ &+ \sup_{u\in[0,\, b]} \int_0^u |(y(u-v)-K)^+ - (y'(u-v)-K)^+|\, dF(v) \\ \leq{}& \rho \int_0^u \upsilon_b[y, y']\, dF_e(v) + \int_0^u \upsilon_b[y, y']\, dF(v) \\ \leq{}& \kappa \upsilon_b[y, y']. \end{aligned}$$

Because $\kappa < 1$, the mapping Ψ is a contraction.

By the contraction mapping theorem (cf. Theorem 3.2 in Hunter and Nachtergaele [17]), Ψ has a unique fixed point x, i.e., $x = \psi(x)$. This implies that x is the unique solution to Equation (50). □

Lemma A.2. *Assume the same condition as in Lemma A.1. Let $x(\cdot) \in \mathbf{D}([0, a], \mathbb{R})$ be the solution to Equation (50) on some interval $[0, a]$ with $F(a) < 1$. If $h(\cdot)$ satisfies the following condition*

$$h(u) = (h(0)\wedge K)[1-G(u)] + (h(0)-K)^+[1-F(u)], \tag{A1}$$

where $h(0) \geq 0$, $F(\cdot)$ is the same probability distribution function as in (50), and $G(\cdot)$ is a probability distribution function, then the function

$$\lambda \int_0^u (x(v)\wedge K)\, dv - (x(u)-K)^+$$

is nondecreasing in u on the interval $[0, a]$.

Proof. To simplify the notation, let $q(u) = (x(u)-K)^+$, $z(u) = x(u)\wedge K$ and

$$b(u) = \lambda \int_0^u z(v)\, dv - q(u) \tag{A2}$$

for all $u \in [0, a]$. We need to show that $b(\cdot)$ is an nondecreasing function on the interval $[0, a]$. It follows from the definition of $F_e(\cdot)$ that $\rho \int_0^u z(u-v)\, dF_e(v) = \lambda \int_0^u z(v)\, dv - \lambda \int_0^u z(v)F(u-v)\, dv$. Plugging it into (50) gives

$$x(u) = h(u) + \lambda \int_0^u z(v)\, dv + \int_0^u q(u-v) dF(v) - \lambda \int_0^u z(v)F(u-v)\, dv.$$

Applying Fubini's Theorem (cf. Theorem 8.4 in Lang [23]) to the last integral in the above, we have

$$\begin{aligned} \lambda \int_0^u z(v)F(u-v)\, dv &= \lambda \int_0^u \int_0^{u-v} z(v)\, dF(x)\, dv \\ &= \lambda \int_0^u \int_0^{u-x} z(v)\, dv\, dF(x). \end{aligned}$$

Thus, we obtain

$$x(u) - \lambda \int_0^u z(v)\,dv = h(u) + \int_0^u \left[q(u-v) - \lambda \int_0^{u-v} z(x)\,dx \right] dF(v).$$

According to the definition of $b(\cdot)$ in (A2), we have

$$b(u) = z(u) - h(u) + \int_0^u b(u-v)\,dF(v). \tag{A3}$$

It now remains to use (A2) and (A3) to argue that $b(\cdot)$ is nondecreasing on the interval $[0, a]$, i.e., for any $u, u' \in [0, a] > 0$ with $u \le u'$, we have $b(u) \le b(u')$. Applying (A3), we have

$$\begin{aligned} b(u') - b(u) &= z(u') - z(u) - [h(u') - h(u)] + \int_0^{u'} b(u'-v)\,dF(v) + \int_0^u b(u-v)\,dF(v) \\ &= z(u') - z(u) - [h(u') - h(u)] + \int_u^{u'} b(u'-v)\,dF(v) + \int_0^u [b(u'-v) - b(u-v)]\,dF(v). \end{aligned}$$

Note that by condition (A1), we have

$$\begin{aligned} -[h(u') - h(u)] &= -(h(0) \wedge K)[G(u) - G(u')] - (h(0) - K)^+ [F(u) - F(u')] \\ &= (h(0) \wedge K)[G(u') - G(u)] - b(0)[F(u') - F(u)], \end{aligned}$$

where the last equation is because of (50) and (A2). Thus,

$$\begin{aligned} b(u') - b(u) &= z(u') - z(u) + (h(0) \wedge K)[G(u') - G(u)] \\ &\quad + \int_u^{u'} [b(u'-v) - b(0)]\,dF(v) + \int_0^u [b(u'-v) - b(u-v)]\,dF(v). \end{aligned} \tag{A4}$$

Because $b \in \mathbf{D}([0, a], \mathbb{R})$, according to Theorem 6.2.2 in the supplement of Whitt [30], it is bounded on the interval $[0, a]$. Let

$$b^* = \inf_{\{(u,u') \in [0,a] \times [0,a]:\, u \le u'\}} b(u') - b(u).$$

If $z(u') < K$, then $q(u') = 0$. Thus, by (A2),

$$b(u') - b(u) = \lambda \int_u^{u'} z(v)\,dv + q(u),$$

which is always nonnegative; if $z(u') = K$, then $z(u') - z(u) \ge 0$. It follows from (A4) that

$$\begin{aligned} b(u') - b(u) &\ge \int_u^{u'} [b(u'-v) - b(0)]\,dF(v) + \int_0^u [b(u'-v) - b(u-v)]\,dF(v) \\ &\ge \int_0^{u'} b^*\,dF(v) = b^* F(u'). \end{aligned}$$

Summarizing both cases, we have

$$b(u') - b(u) \ge \min(0, b^* F(u'))$$

for all $u, u' \in [0, a] > 0$ with $u \le u'$. Suppose that $b^* < 0$. Taking the infimum on both sides over the set $\{(u, u') \in [0, a] \times [0, a]:\, u \le u'\}$ gives $b^* \ge F(a) b^*$. This implies that $[1 - F(a)] b^* \ge 0$. Because $F(a) < 1$, it contradicts to that $b^* < 0$. Thus, we must have $b^* \ge 0$, which implies that $b(\cdot)$ is nondecreasing on $[0, a]$. □

Appendix B. Glivenko-Cantelli estimate. The Glivenko-Cantelli estimate, cf. Lemma B.1 here, was used in several places in this paper. A very similar result was proved in Lemma 5.1 (Gromoll et al. [15]). The differences only stay at the technical level. For completeness, the proof which follows the one in Gromoll et al. [15] is provided here.

For any r, consider the sequence of i.i.d random variables $\{v_i^r\}_{i=-\infty}^{\infty}$ with law ν^r. In our setting, those v_i^r's with $i \ge 1$ correspond to the service requirement of the arriving jobs in the rth system and those with $i \le 0$ correspond to the service requirement of initial jobs in the rth system. Assume that

$$\nu^r \to \nu \quad \text{as } r \to \infty.$$

For any $n \in \mathbb{Z}$ and $l \in \mathbb{R}_+$, define

$$\bar{\eta}^r(n,l) = \frac{1}{r} \sum_{i=n+1}^{n+\lfloor rl \rfloor} \delta_{v_i^r}. \tag{B1}$$

Let $\mathscr{C} = \{[y,\infty)\colon y \in \mathbb{R}_+\} \cup \{(y,\infty)\colon y \in \mathbb{R}_+\}$ and

$$\mathscr{V} = \{1_C\colon C \in \mathscr{C}\}.$$

The Skorohod representation theorem implies the existence of $\mathbb{R}_+$-valued random variables $Y^r \sim \nu^r$ and $Y \sim \nu$ such that $Y^r \to Y$ almost surely. Thus, there exists an $\mathbb{R}_+$-valued random variable $\bar{Y}$ such that

$$\bar{Y} = \sup_{r \in \mathbb{R}_+} Y^r, \quad \text{almost surely.} \tag{B2}$$

Let $\bar{\nu}$ be the law of $\bar{Y}$. The space $L_2(\bar{\nu})$ of all Borel measurable functions $f\colon \mathbb{R}_+ \to \mathbb{R}$ equipped with the $L_2(\bar{\nu})$-norm $\|f\|_{\bar{\nu},2} = \langle |f|^2, \bar{\nu}\rangle$ contains a continuous, increasing, and unbounded function $\bar{f}$ such that $\bar{f} \geq 1$ and

$$\mathbb{E}[\bar{f}(\bar{Y})^2] = \langle \bar{f}^2, \bar{\nu}\rangle < \infty. \tag{B3}$$

Because $1_C \leq \bar{f}$ for all $C \in \mathscr{C}$, we call $\bar{f}$ an envelope function for $\mathscr{V}$. Finally, denote $\bar{\mathscr{V}} = \mathscr{V} \cup \{\bar{f}\}$. The objective of this section is to obtain the following *Glivenko-Cantelli Estimate* for $\bar{\eta}^r(n,l)$.

Lemma B.1. *Assume that* $\mathbf{d}[\nu^r, \nu] \to 0$ *as* $r \to \infty$, *where* ν *is a probability measure. Fix constants* M_0, M_1, $L > 0$. *For all* $\epsilon, \eta > 0$,

$$\limsup_{r\to\infty} \mathbb{P}^r\left(\max_{-rM_0 < n < rM_1} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathscr{V}}} |\langle f, \bar{\eta}^r(n,l)\rangle - l\langle f, \nu^r\rangle| > \epsilon \right) < \eta. \tag{B4}$$

To prove the result, we introduce some notions from empirical process theory. Our primary references are Gromoll et al. [15] and van der Vaart and Wellner [29]. A collection $\mathscr{C}$ of subsets of $\mathbb{R}_+$ shatters an n-point subset $\{x_1, \ldots, x_n\} \subset \mathbb{R}_+$ if the collection $\{\mathscr{C} \cap \{x_1, \ldots, x_n\}\colon C \in \mathscr{C}\}$ has cardinality 2^n. In this case, we say that $\mathscr{C}$ picks out all subsets of $\{x_1, \ldots, x_n\}$. The *Vapnik-Červonenkis index (VC-index)* of $\mathscr{C}$ is

$$V_{\mathscr{C}} = \min\{n\colon \mathscr{C} \text{ shatters no } n\text{-point subset}\},$$

where the minimum of the empty set equals infinity. The collection $\mathscr{C}$ is a *Vapnik-Červonenkis class* (*VC-class*) if it has finite VC-index. In our case, $\mathscr{C} = \{[x,\infty)$ and $(x,\infty)\colon x \in \mathbb{R}_+\}$. It is easy to see that $\mathscr{C}$ shatters no two-point subset, so it has VC-index bounded above by two. Thus, $\mathscr{C}$ is a VC-class.

VC-classes satisfy a very useful entropy bound. Let Γ be the set of all Borel probability measures γ on $\mathbb{R}_+$. For all $\gamma \in \Gamma$, denote $L_1(\gamma)$ the space of all Borel measurable functions $f\colon \mathbb{R}_+ \to \mathbb{R}$ equipped with $L_1(\gamma)$-norm

$$\|f\|_{\gamma,1} = \langle |f|, \gamma\rangle.$$

For any $f \in L_1(\gamma)$, let $B_\gamma(f,\epsilon) = \{g \in \mathscr{V}\colon \|g - f\|_{\gamma,1} < \epsilon\}$ denote the $L_1(\gamma)$-ball in $L_1(\gamma)$, centered at f with radius ϵ. For a family of functions $\mathscr{V}$, $N(\epsilon, \mathscr{V}, L_1(\gamma))$ is the smallest number of balls $B_\gamma(f,\epsilon)$ needed to cover $\mathscr{V}$. Because $\mathscr{V}$ is the set of index functions over a VC-class $\mathscr{C}$,

$$\sup_{\gamma\in\Gamma} \log N(\epsilon\|\bar{f}\|_{\gamma,1}, \mathscr{V}, L_1(\gamma)) < \infty; \tag{B5}$$

see Theorem 2.6.4 in van der Vaart and Wellner [29].

Proof of Lemma B.1. Define

$$\bar{\eta}^r(l) = \frac{1}{r} \sum_{i=\lfloor -rM_0 \rfloor + 1}^{\lfloor -rM_0 \rfloor + \lfloor rl \rfloor} \delta_{v_i^r}.$$

By (B1), it suffices to show that

$$\limsup_{r\to\infty} \mathbb{P}^r\left(\sup_{l\in[0,L']} \sup_{f\in\bar{\mathscr{V}}} |\langle f, \bar{\eta}^r(l)\rangle - l\langle f, \nu^r\rangle| > \epsilon/2 \right) < \eta, \tag{B6}$$

where $L' = L + M_0 + M_1$.

We now apply Theorem 2.8.1 in van der Vaart and Wellner [29] to show (B6). Observe that for all $n \in \mathbb{N}$ and $(e_1, \ldots, e_n) \in \mathbb{R}^n$, the function

$$(x_1, \ldots, x_n) \to \sup_{f \in \bar{\mathscr{V}}} \sum_{i=1}^{n} e_i f(x_i)$$

is measurable on the completion of $(\bar{\mathbb{R}}_+, \mathscr{B}, \nu^r)^n$, for all $r \in \mathbb{R}_+$. Thus, $\bar{\mathscr{V}}$ is a ν^r-measurable class for all $r \in \mathbb{R}_+$; see Definition 2.3.3 in van der Vaart and Wellner [29]. Moreover, $\bar{\mathscr{V}}$ is uniformly bounded above by the envelope function $\bar{f}$ and

$$\lim_{M \to \infty} \sup_{r \in \mathbb{R}_+} \langle \bar{f} 1_{\{\bar{f} > M\}}, \nu^r \rangle = 0 \tag{B7}$$

by Markov's inequality, (B2), and (B3). Last, $\bar{\mathscr{V}}$ satisfies the finite entropy bound (B5) because $N(\epsilon, \bar{\mathscr{V}}, L_1(\gamma)) \le N(\epsilon, \mathscr{V}, L_1(\gamma)) + 1$ and because $\mathscr{C}$ is a VC-class. These three observations imply that the assumptions of Theorem 2.8.1 in van der Vaart and Wellner [29] are satisfied. Consequently, $\bar{\mathscr{V}}$ is *Glivenko-Cantelli*, uniformly in r. That is, for every $\delta > 0$, there exists an n_δ such that

$$\limsup_{r \to \infty} \mathbb{P}^r \left(\sup_{m \ge n_\delta} \sup_{f \in \bar{\mathscr{V}}} \left| \frac{1}{m} \sum_{-\lfloor rM_0 \rfloor + 1}^{-\lfloor rM_0 \rfloor + m} f(v_i^r) - \langle f, \nu^r \rangle \right| > \delta \right) < \delta. \tag{B8}$$

Note that the probability on the left-hand side of (B6) can be upper bounded by

$$\mathbb{P}^r \left(\sup_{l \in [0, L']} \sup_{f \in \bar{\mathscr{V}}} \frac{\lfloor rl \rfloor}{r} \left| \frac{1}{\lfloor rl \rfloor} \sum_{-\lfloor rM_0 \rfloor + 1}^{-\lfloor rM_0 \rfloor + \lfloor rl \rfloor} f(v_i^r) - \langle f, \nu^r \rangle \right| > \epsilon/4 \right) + \mathbb{P}^r \left(\frac{1}{r} \sup_{f \in \bar{\mathscr{V}}} \langle f, \nu^r \rangle > \frac{\epsilon}{4} \right).$$

By (B2) and (B3), the second term in the above vanishes as $r \to \infty$. The first term can be upper bounded by

$$\begin{aligned} &\mathbb{P}^r \left(\frac{n_\delta}{r} \sup_{m \in [0, n_\delta]} \sup_{f \in \bar{\mathscr{V}}} \left| \frac{1}{m} \sum_{-\lfloor rM_0 \rfloor + 1}^{-\lfloor rM_0 \rfloor + m} f(v_i^r) - \langle f, \nu^r \rangle \right| > \epsilon/4 \right) \\ &\quad + \mathbb{P}^r \left(L' \sup_{m \in [n_\delta, L'r]} \sup_{f \in \bar{\mathscr{V}}} \left| \frac{1}{m} \sum_{-\lfloor rM_0 \rfloor + 1}^{-\lfloor rM_0 \rfloor + m} f(v_i^r) - \langle f, \nu^r \rangle \right| > \epsilon/4 \right). \end{aligned} \tag{B9}$$

To see this, one can replace m by $\lfloor rl \rfloor$ and divide the interval $[0, L']$ into $[0, n_\delta/r]$ and $[n_\delta/r, L']$. Denote

$$X(f) = \sup_{m \in [0, n_\delta]} \left| \frac{1}{m} \sum_{-\lfloor rM_0 \rfloor + 1}^{-\lfloor rM_0 \rfloor + m} f(v_i^r) - \langle f, \nu^r \rangle \right|.$$

When $f \in \mathscr{V}$, it is clear that $X(f) \le 2$. By (B2) and (B3), $X(\bar{f})$ is a random variable with finite mean and variance. Thus, there exists a constant M_3 such that

$$\mathbb{P}^r \left(\sup_{f \in \bar{\mathscr{V}}} X(f) > M_3 \right) < \eta/2.$$

The first term in (B9) is bounded by $\eta/2$ for all $r \ge 4M_3 n_\delta/\epsilon$. According to (B8), the lim sup of the second term in (B9) will be bounded by $\eta/2$ if we choose $\delta = \min(\epsilon/(4L'), \eta/2)$. □

Acknowledgements. This research is supported in part by National Science Foundation Grants CMMI-0727400 and CNS-0718701 and by an IBM Faculty Award. The authors thank two anonymous referees for significantly improving the paper.

References

[1] Avi-Itzhak, B., S. Halfin. 1988. Expected response times in a non-symmetric time sharing queue with a limited number of service positions. *Proc. 12th Internat. Teletraffic Congress, Torino, Italy*.
[2] Billingsley, P. 1995. *Probability and Measure*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. John Wiley & Sons Inc., New York.
[3] Billingsley, P. 1999. *Convergence of Probability Measures*, 2nd ed. *Wiley Series in Probability and Statistics: Probability and Statistics*. John Wiley & Sons Inc., New York.
[4] Blake, R. 1982. Optimal control of thrashing. *Proc. ACM SIGMETRICS Conf. Measurements Modeling Comput. Systems, Seattle*.
[5] Chung, K. L. 2001. *A Course in Probability Theory*, 3rd ed. Academic Press Inc., San Diego.

[6] Dai, J. G. 1995a. On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **5**(1) 49–77.
[7] Dai, J. G. 1995b. Stability of open multiclass queueing networks via fluid models. *Proc. IMA Workshop Stochastic Networks*, Springer-Verlag, New York.
[8] Denning, P. J., K. C. Kahn, J. Leroudier, D. Potier, R. Suri. 1976. Optimal multiprogramming. *Acta Informatica* **7** 197–216.
[9] Doytchinov, B., J. Lehoczky, S. Shreve. 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* **11**(2) 332–378.
[10] Elnikety, S., E. Nahum, J. Tracy, W. Zwaenepoel. 2004. A method for transparent admission control and request scheduling in e-commerce websites. *World Wide Web Conf., New York*.
[11] Ethier, S. N., T. G. Kurtz. 1986. *Markov Processes*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. John Wiley & Sons Inc., New York.
[12] Gromoll, H. C. 2004. Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.* **14**(2) 555–611.
[13] Gromoll, H. C., Ł. Kruk. 2007. Heavy traffic limit for a processor sharing queue with soft deadlines. *Ann. Appl. Probab.* **17**(3) 1049–1101.
[14] Gromoll, H. C., A. L. Puha, R. J. Williams. 2002. The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.* **12**(3) 797–859.
[15] Gromoll, H. C., P. Robert, B. Zwart. 2008. Fluid limits for processor sharing queues with impatience. *Math. Oper. Res.* **33**(2) 375–402.
[16] Heiss, H.-U., R. Wagner. 1991. Adaptive load control in transaction processing systems. *Proc. 17th Internat. Conf. Large Data Bases, Barcelona, Spain*.
[17] Hunter, J. K., B. Nachtergaele. 2001. *Applied Analysis*. World Scientific Publishing Co. Inc., River Edge, NJ.
[18] Jean-Marie, A., P. Robert. 1994. On the transient behavior of the processor sharing queue. *Queueing Systems Theory Appl.* **17**(1–2) 129–136.
[19] Kallenberg, O. 1986. *Random Measures*, 4th ed. Akademie-Verlag, Berlin.
[20] Kamra, A., V. Misra, E. M. Nahum. 2004. Yaksha: A self-tuning controller for managing the performance of 3-tiered web sites. *12th IEEE Internat. Workshop Quality Service, Montréal*.
[21] Kaspi, H., K. Ramanan. 2007. Law of large numbers limits for many-server queues. Working paper.
[22] Kleinrock, L. 1976. *Queueing Systems*. Vol. II, *Computer Applications*. Wiley-Interscience, New York.
[23] Lang, S. 1983. *Real Analysis*, 2nd ed. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA.
[24] Nuyens, M., W. van der Weij. 2007. The limited processor sharing queue. Technical report, Centrum voor Wiskunde en Informatica, Amsterdam.
[25] Puha, A. L., R. J. Williams. 2004. Invariant states and rates of convergence for a critical fluid model of a processor sharing queue. *Ann. Appl. Probab.* **14**(2) 517–554.
[26] Puha, A. L., A. L. Stolyar, R. J. Williams. 2006. The fluid limit of an overloaded processor sharing queue. *Math. Oper. Res.* **31**(2) 316–350.
[27] Ritchie, D. M., K. Thompson. 1974. The Unix time-sharing system. *J. ACM* **17**(7) 365–375.
[28] Schroeder, B., M. Harchol-Balter, A. Iyengar, E. Nahum, A. Wierman. 2006. How to determine a good multi-programming level for external scheduling. *Proc. 22nd Internat. Conf. Data Engrg., Atlanta*.
[29] van der Vaart, A. W., J. A. Wellner. 1996. *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. Springer-Verlag, New York.
[30] Whitt, W. 2002. *Stochastic-Process Limits*. Springer-Verlag, New York.
[31] Zhang, F., L. Lipsky. 2006. Modelling restricted processor sharing. *Proc. Internat. Conf. Parallel Distributed Processing Techniques Appl.* (*PDPTA06*), *Las Vegas, NV*.
[32] Zhang, F., L. Lipsky. 2007. An analytical model for computer systems with non-exponential service times and memory thrashing overhead. *Proc. Internat. Conf. Parallel Distributed Processing Techniques Appl.* (*PDPTA07*), *Las Vegas, NV.*
[33] Zhang, J., B. Zwart. 2008. Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Systems Theory Appl.* **60**(3–4) 227–246.
[34] Zhang, J., J. G. Dai, B. Zwart. 2007. Diffusion limits of limited processor sharing queues. Technical report, Georgia Institute of Technology, Atlanta. http://www.isye.gatech.edu/~jzhang/research/lps-ht.pdf.