The Annals of Applied Probability
2011, Vol. 21, No. 2, 745–799
DOI: 10.1214/10-AAP709

DIFFUSION LIMITS OF LIMITED PROCESSOR SHARING QUEUES[1]

By Jiheng Zhang, J. G. Dai and Bert Zwart[2]

Hong Kong University of Science and Technology, Georgia Institute of Technology and CWI

We consider a processor sharing queue where the number of jobs served at any time is limited to K, with the excess jobs waiting in a buffer. We use random counting measures on the positive axis to model this system. The limit of this measure-valued process is obtained under diffusion scaling and heavy traffic conditions. As a consequence, the limit of the system size process is proved to be a piece-wise reflected Brownian motion.

1. Introduction. This paper is concerned with developing a diffusion approximation for a *limited processor sharing* (LPS) queue, which consists of a single server and an infinite capacity buffer. In such a system, the server can serve up to $K \geq 1$ jobs simultaneously, equally distributing its attention to each of them. In other words, 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 customer from the buffer, if there is one. We assume that jobs arrive according to a general arrival process, and the job sizes are independent of each other and identically distributed.

Received December 2007; revised April 2010.

[1]Supported in part by NSF Grants CMMI-0727400, CMMI-0825840, CNS-0718701 and DMS-08-05979 and by two IBM Faculty awards and a VIDI grant from NWO.

[2]This author is also affiliated with EURANDOM, VU University Amsterdam, and Georgia Institute of Technology.

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

Key words and phrases. Limited processor sharing, heavy traffic, diffusion approximation, state-space collapse, measure valued process.

Note that letting $K = \infty$ makes the system a standard *processor sharing* (PS) queue, which has been the focal point of significant recent research activity. The PS discipline can be viewed as an idealization of time-sharing protocols in computer systems, as described in [20] and [23]. 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 due to switching, and hence reduce overall performance. This point has already been observed in early studies of operating systems [5, 8], as well as in more recent Web server design papers [10, 18] and database implementation papers [16, 24]. So in the modeling of many computer and communication systems, a sharing limit is normally imposed, which results in the LPS model.

Despite the numerous applications, there are only a few studies on the LPS queue. Avi-Itzhak and Halfin [2] propose an approximation for the mean response time assuming Poisson arrivals. A computational analysis based on matrix geometric methods is performed in Zhang and Lipsky [27, 28]. Some stochastic ordering results are derived in Nuyens and van de Weij [21]. Recently, Zhang, Dai and Zwart [29] have developed a fluid approximation for the LPS queue using the framework of measure-valued processes. As a continuation of [29], the present study investigates a diffusion approximation for the LPS queue in the heavy traffic regime.

In our model, the system consists of a server for serving jobs and a buffer for holding the waiting jobs. We model the LPS queue by means of a measure-valued process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$. Each component of the process takes values in the space of finite, nonnegative Borel measures on $\mathbb{R}_+ = [0, \infty)$. For each $t \geq 0$, $\mathcal{Z}(t)$ puts unit mass at the residual job size of each job in the server at time $t \geq 0$, and $\mathcal{Q}(t)$ puts unit mass at the job size of each job in the buffer at time $t \geq 0$. The main insight of our approach is to design the stochastic dynamic equations (2.5) and (2.6) using the measure valued process $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot))$, which can describe the evolution of the system. Our asymptotic regime is when the sharing limit K is large and the queue is critically loaded. Following the standard practice in the literature, we consider a sequence of queues indexed by $r \in \mathbb{R}_+$. We assume that $\lim_{r\to\infty} K^r/r = K > 0$ and the traffic intensity goes to the critical value 1 as in (2.23), where K^r is the sharing limit in the rth queue. (Superscript r indicates a quantity that is associated with the rth queue.)

We are interested in the limit of the diffusion scaled process

$$\left(\frac{1}{r}\mathcal{Q}^r(r^2\cdot), \frac{1}{r}\mathcal{Z}^r(r^2\cdot)\right)$$

as r goes to infinity. As shown in Williams [26], a key step to obtain a diffusion limit in heavy traffic is to establish a *state-space collapse* (SSC) result. In our setting, the SSC means that the diffusion-scaled measure-valued process, which is an infinite-dimensional object, is close to a deterministic function of the diffusion-scaled, one-dimensional workload process. (See Definition 2.1 for the lifting map to define the function.) The workload process is invariant under any nonidling service policy, and its diffusion limit is a one-dimensional reflected Brownian motion (RBM). The main result of this paper (Theorem 2.1) is that our measure-valued diffusion limit is a deterministic function of the one-dimensional RBM. As a corollary, the diffusion-scaled system size process converges in distribution to a piecewise linear RBM.

Most of this paper is devoted to proving the SSC result for each fixed time $T>0$ (Theorem 2.2). Our proof strategy is analogous to the modular approach proposed in Bramson [6] and Williams [26]. For our sequence of systems, we define a critically loaded measure-valued fluid model. The fluid model in this paper is the same LPS fluid model developed in [29], specialized for the critically loaded case. We show that our fluid model exhibits an SSC: each fluid model solution converges to an equilibrium state in some uniform sense, and each equilibrium state has an SSC.

We adopt Bramson's framework in [6] to translate the fluid model SSC result into the diffusion-scaled SSC result. The diffusion scaled process on the interval $[0,T]$ corresponds to unscaled process on the interval $[0,r^2T]$. Fix a constant $L>1$, the interval $[0,r^2T]$ is covered by the $\lfloor rT \rfloor + 1$ overlapping intervals

$$[rm, rm+rL], \qquad m=0,1,\ldots,\lfloor rT \rfloor.$$

On each of these intervals, the diffusion scaled process can be viewed as a shifted, fluid-scaled process defined by

$$\left(\frac{1}{r}\mathcal{Q}^r(rm+rt), \frac{1}{r}\mathcal{Z}^r(rm+rt)\right), \qquad 0 \le t \le L.$$

To carry out the translation, we need to show that (a) each limit from the family of shifted, fluid-scaled processes is a solution to the fluid model (such a limit is called a fluid limit in this paper, also known as a "cluster point" in [6]); (b) the set of fluid limits is "rich": with large probability, each shifted, fluid-scaled measure-valued process is close to some fluid limit. A major step to proving (a) and (b) is to show, with large probability, the precompactness of the shifted fluid scaled processes (see Theorem 4.1 for details). Since m ranges from 0 to $\lfloor rT \rfloor$, this involves a substantial refinement of the arguments in [29], where the case $m=0$ is treated.

Establishing SSC for the fluid model requires a study of equilibrium states for the LPS fluid model developed in [29]. In Section 3, we characterize the set of equilibrium states, and show that each fluid model solution with initial condition belonging to a compact set converges uniformly to its equilibrium state. The counterpart of this study for standard PS queues has been carried out in [22]. Standard PS queues are relatively tractable since their fluid models can be related (by means of a time-change) to a renewal equation. This is not the case for LPS systems (as explained in [29]), so a different approach is necessary. The idea behind the proof of the uniform convergence is to carefully track the total mass of the fluid model, and determine whether it is eventually bigger, smaller or equal to the sharing limit K. Several insights (explained in Sections 3.2 and 3.3) lead us to apply a uniform version of the renewal theorem which is new to the best of our knowledge. This version is given in Appendix B.1.

The framework of measure-valued process has been successfully applied to study models where multiple jobs are processed at the same time. The main idea is to use a sufficiently detailed state descriptor to adequately describe the system. A sequence of papers, Gromoll, Puha and Williams [14], Puha and Williams [22] and Gromoll [12], has successfully established the fluid and diffusion approximations for PS queues using measure-valued processes. More recently, the framework of measure-valued process has been further developed by Gromoll and Kruk [13] and Gromoll, Robert and Zwart [15] in the study of queues with deadlines/impatience. Doytchinov, Lehoczky and Shreve [9] applied a similar framework to study the earliest deadline first discipline. This framework is also applied by Kaspi and Ramanan [19] on many-server queues. The results in the present paper can be seen as an extension of the results in the papers [12, 22], which carry out a similar program for the standard PS queue.

This paper is organized as follows. A model description and an overview of the main results is given in Section 2. Section 3 investigates the convergence of each fluid model solution to its equilibrium state. Precompactness of the family of shifted fluid scaled processes is established in Section 4. Section 5 uses the precompactness to show the "richness" of the fluid limits, and then concludes with a proof of state-space collapse. Several additional useful results, such as a uniform version of the renewal theorem, and a useful bound for the Prohorov metric are developed in Appendices B and C.

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 $\mathbf{M}$ denote the set of all nonnegative finite Borel measures on $[0,\infty)$. To simplify the notation, let us take the convention that for any Borel set $A \subseteq \mathbb{R}$, $\nu(A \cap (-\infty, 0)) = 0$ for any $\nu \in \mathbf{M}$. For $\nu_1, \nu_2 \in \mathbf{M}$, the Prohorov metric is defined to be

$$\mathbf{d}[\nu_1, \nu_2] = \inf\{\varepsilon > 0 : \nu_1(A) \le \nu_2(A^\varepsilon) + \varepsilon \text{ and}$$

$$\nu_2(A) \le \nu_1(A^\varepsilon) + \varepsilon \text{ for all Borel set } A \subseteq \mathbb{R}\},$$

where $A^\varepsilon = \{b \in \mathbb{R} : \inf_{a\in A}|a-b| < \varepsilon\}$. This same metric was defined and used in Gromoll and Kruk [13]; they showed that the space $\mathbf{M}$ is complete and separable under the metric. For any Borel measurable function $g : \mathbb{R}_+ \to \mathbb{R}$, the integration of this function with respect to the measure $\nu \in \mathbf{M}$, $\int_{\mathbb{R}_+} g(x)\nu(dx)$, is denoted by $\langle g, \nu\rangle$.

Let $\mathbf{M} \times \mathbf{M}$ 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

$$\upsilon_T[g, g'] = \sup_{0\le t\le T} \pi[g(t), g'(t)]. \tag{1.1}$$

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

$$\varrho_T[g, g'] = \inf_{f\in\Lambda_T} (\|f\|_T^\circ \vee \upsilon_T[g, g' \circ f]), \tag{1.2}$$

where $g \circ f(\cdot) = g(f(\cdot))$ 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

$$\varrho[g, g'] = \int_0^\infty e^{-T} (\varrho_T[g, g'] \wedge 1)\, dT. \tag{1.3}$$

By "convergence in the space $\mathbf{D}$," we mean the convergence under the Skorohod J_1 topology, which is induced by the Skorohod J_1 metric [11].

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

2. Models and main results. In this section, we first introduce the mathematical model. We then present the main results of this paper. Following this, is an outline of our proof.

2.1. *The limited processor sharing queue.* 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, the number of jobs in service, and the total number of jobs in the system at time t, respectively. Thus,

$$X(t) = Q(t) + Z(t) \qquad \text{for } t \geq 0. \tag{2.1}$$

We adopt the convention that $Q(\cdot), Z(\cdot)$ and $X(\cdot)$ are right continuous. The system is allowed to be nonempty initially, that is, $X(0) > 0$. We index jobs by $i = -X(0)+1, -X(0)+2, \dots, 0, 1, \dots$. The first $X(0)$ jobs are initially in the system, with jobs $i = -X(0)+1, \dots, -Q(0)$ in service and jobs $i = -Q(0)+1, \dots, 0$ waiting in the buffer. Jobs arrived after time 0 are indexed by $i = 1, 2, \dots,$ according to the order of arrival. When a batch arrival occurs, an arbitrary rule is used to break the tie for the arrivals in the batch. The service policy in this model is FCFS. Let $E(t)$ denote the number of jobs that arrive at the buffer during time interval $(0,t]$, for all $t \geq 0$. Our arrival process $\{E(t), t \geq 0\}$ is assumed to be general, as long as it satisfies a functional central limit theorem [see (2.14)]. 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 $i \leq -Q(0)$. Let

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

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 $F(\cdot)$. Denote ν the probability measure associated with the distribution function $F(\cdot)$. For jobs with index $-X(0) < i \leq -Q(0)$, that is, the first $Z(0)$ jobs that are initially in service, we use $\tilde{v}_i$ to denote the remaining job size of the job. The sequence $\{\tilde{v}_i\}_{i=-\infty}^{0}$ is allowed to be general. 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.

Now we introduce a measure-valued state descriptor $(\mathcal{Q}(\cdot), \mathcal{Z}(\cdot)) \in \mathbf{M} \times \mathbf{M}$, which describes the evolution of the system with given initial conditions and stochastic primitives. For any Borel set $A \subset [0, \infty)$, $\mathcal{Q}(t)(A)$ denotes the total number of jobs in the 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 A. Since no job can be in service with residual job size 0, $\mathcal{Z}(t)(\{0\}) = 0$ for all $t \geq 0$. 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

$$S(t) = \int_0^t \psi(Z(\tau))\, d\tau, \tag{2.2}$$

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

$$B(t) = E(t) - Q(t). \tag{2.3}$$

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

$$B(s,t) = B(t) - B(s) \tag{2.4}$$

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

$$\mathcal{Q}(t)(A) = \sum_{i=B(t)+1}^{E(t)} \delta_{v_i}(A), \qquad A \subset [0,\infty), \tag{2.5}$$

$$\begin{aligned} \mathcal{Z}(t)(A) &= \sum_{i=-X(0)+1}^{-Q(0)} \delta_{\tilde{v}_i}(A + S(t)) \\ &\quad + \sum_{i=-Q(0)+1}^{B(t)} \delta_{v_i}(A + S(\tau_i, t)), \qquad A \subset (0,\infty), \end{aligned} \tag{2.6}$$

with $\mathcal{Z}(t)(\{0\}) = 0$, where $\delta_a(A)$ denotes the Dirac measure of point a on $\mathbb{R}$ and $A + y = \{a + y : a \in A\}$. Due to the LPS policy, the sharing limit K must be enforced at any time t,

$$Q(t) = (X(t) - K)^+, \tag{2.7}$$

$$Z(t) = (X(t) \wedge K). \tag{2.8}$$

We call (2.5) and (2.6) the *stochastic dynamic equations* and (2.7) and (2.8) the policy constraints.

For $t \geq 0$, 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 the state descriptor $(\mathcal{Q}, \mathcal{Z})$, we can recover the workload $W(t)$ at time $t \geq 0$ by

$$W(t) = \langle \chi, \mathcal{Q}(t) + \mathcal{Z}(t) \rangle, \tag{2.9}$$

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

2.2. *Main results.* Consider a sequence of limited processor sharing queues indexed by r, where r increases to ∞ through a sequence in $(0,\infty)$. Each queue is defined in the same way as in Section 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 *diffusion* scaling, which is defined by

$$\hat{\mathcal{Q}}^r(t) = \frac{1}{r}\mathcal{Q}^r(r^2 t), \qquad \hat{\mathcal{Z}}^r(t) = \frac{1}{r}\mathcal{Z}^r(r^2 t), \tag{2.10}$$

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

$$\hat{Q}^r(t) = \langle 1, \hat{\mathcal{Q}}^r(t)\rangle = \frac{1}{r}Q^r(r^2 t), \tag{2.11}$$

$$\hat{Z}^r(t) = \langle 1, \hat{\mathcal{Z}}^r(t)\rangle = \frac{1}{r}Z^r(r^2 t), \tag{2.12}$$

$$\hat{W}^r(t) = \langle \chi, \hat{\mathcal{Q}}^r(t) + \hat{\mathcal{Z}}^r(t)\rangle = \frac{1}{r}W^r(r^2 t), \tag{2.13}$$

for all $t \ge 0$.

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

$$\frac{E^r(r^2\cdot) - \lambda^r r^2\cdot}{r} \Rightarrow E^*(\cdot) \qquad \text{as } r \to \infty, \tag{2.14}$$

for some sequence $\{\lambda^r\}$ that satisfies

$$\lim_{r\to\infty} \lambda^r = \lambda > 0, \tag{2.15}$$

and $E^*(\cdot)$ is a Brownian motion with drift 0 and variance λc_a^2. And the probability measure ν^r of job sizes in the rth system satisfies that as $r \to \infty$

$$\mathbf{d}[\nu^r, \nu] \to 0, \tag{2.16}$$

$$\lim_{N\to\infty} \sup_r \int_{[N,\infty)} x^{4+2p}\nu^r(dx) \to 0 \qquad \text{for some } p > 0, \tag{2.17}$$

where the probability measure ν satisfies

$$\nu \text{ has no atoms.} \tag{2.18}$$

Assumption (2.17) is stronger than the "two-plus-epsilon moment" assumption needed for a functional central limit theorem. The stronger assumption is used in a separate part of our analysis to estimate moments of the shifted fluid scaled state descriptors (see Lemma D.1, and its application in Lemma 4.3). The extra moment assumption also appears in [12, 13]. Since the space is scaled by r in the diffusion scaling, the sharing limit should be scaled accordingly:

$$\lim_{r\to\infty} K^r/r \to K > 0. \tag{2.19}$$

Also, the following initial condition will be assumed:

$$(\hat{\mathcal{Q}}^r(0), \hat{\mathcal{Z}}^r(0)) \Rightarrow (\xi^*, \mu^*), \tag{2.20}$$

$$\langle \chi^{1+p}, \hat{\mathcal{Q}}^r(0) + \hat{\mathcal{Z}}^r(0)\rangle \Rightarrow \langle \chi^{1+p}, \xi^* + \mu^*\rangle, \tag{2.21}$$

as $r \to \infty$, where p is the same as in (2.17), $(\xi^*, \mu^*) \in \mathbf{M} \times \mathbf{M}$ and

$$\mu^* \text{ has no atoms.} \tag{2.22}$$

Define the traffic intensity of the rth stochastic system by $\rho^r = \lambda^r\langle \chi, \nu^r\rangle$. We need the following heavy traffic condition:

$$\lim_{r\to\infty} r(1-\rho^r) = \theta > 0. \tag{2.23}$$

Let $\beta = \langle \chi, \nu\rangle$ be the mean and $c_s^2 = \frac{\langle \chi^2, \nu\rangle - \beta^2}{\beta^2}$ be the squared coefficient of variation (SCV) of the job size distribution ν. The following proposition is a well-known heavy traffic approximation for the workload process of a single queue operated under a nonidling policy. Readers are referred to [12] for a proof.

PROPOSITION 2.1. *Assume (2.14)–(2.17), (2.20), (2.21) and (2.23). The sequence of diffusion scaled workload process*

$$\hat{W}^r(\cdot) \Rightarrow W^*(\cdot) \qquad \text{as } r \to \infty,$$

where $W^*(\cdot)$ *is a reflected Brownian motion with drift* $-\theta$, *variance* $\beta(c_a^2 + c_s^2)$ *and initial value* $w^* = \langle \chi, \xi^* + \mu^*\rangle$.

Since the LPS is also a nonidling service policy, the above result on the workload process is still true for our model. However, it remains an open question about the job size process $X(\cdot)$ and many other performance processes as introduced in Section 2.1. Our main result establishes the diffusion limit for the measure-valued processes (Theorem 2.1), from which the diffusion limit of queue length process follows directly (Corollary 2.1).

Denote

$$\beta_e = \langle \chi, \nu_e\rangle,$$

where ν_e is the *equilibrium* measure of ν, that is, $\nu_e([0,x]) = \frac{1}{\beta}\int_0^x \nu((y,\infty))\,dy$ for all $x \geq 0$. We have the following definition.

DEFINITION 2.1. Let $\Delta_{K,\nu}:\mathbb{R}_+ \to \mathbf{M}\times\mathbf{M}$ be the lifting map associated with the probability measure ν and constant K given by

$$\Delta_{K,\nu}w = \left(\frac{(w-K\beta_e)^+}{\beta}\nu, \frac{w\wedge K\beta_e}{\beta_e}\nu_e\right) \qquad \text{for } w\in\mathbb{R}_+.$$

Note that $\Delta_{K,\nu}$ maps the workload, which is in $\mathbb{R}_+$, to a measure-valued state, which is in $\mathbf{M}\times\mathbf{M}$. The intuition is that the remaining job sizes of those in service have the probability measure ν_e, which is the equilibrium distribution of the job size distribution ν. The total workload embodied in jobs that are in service is $w\wedge K\beta_e$ because at most K jobs are allowed in service. Dividing it by β_e gives the number of jobs in service. The remaining workload $(w-K\beta_e)^+ = w-(w\wedge K\beta_e)$ resides in the buffer, where each job size follows the probability measure ν. Dividing this amount by β, the mean of ν, gives the number of jobs in buffer.

Our main result requires that the limit (ξ^*,μ^*) in (2.20) satisfies

$$(\xi^*,\mu^*) = \Delta_{K,\nu}w^*. \tag{2.24}$$

A similar condition on the initial state is also imposed for proving a diffusion limit in queueing networks [6, 26]. When (2.24) does not hold, it is likely that some delayed versions of theorems are still true as in Theorem 3 of Bramson [6]; however, we will not pursue such generalization in this paper.

THEOREM 2.1. *Assume (2.14)–(2.23). The sequence of diffusion scaled state descriptors*

$$(\hat{\mathcal{Q}}^r(\cdot),\hat{\mathcal{Z}}^r(\cdot)) \Rightarrow \Delta_{K,\nu}W^*(\cdot) \qquad \text{as } r\to\infty,$$

where $W^*(\cdot)$ is the reflected Brownian motion in Proposition 2.1.

COROLLARY 2.1 (Piecewise reflected Brownian motion). *Assume (2.14)–(2.23). The sequence of diffusion scaled system size process*

$$\hat{X}^r(\cdot) = \langle 1, \hat{\mathcal{Q}}^r(\cdot)+\hat{\mathcal{Z}}^r(\cdot)\rangle$$

converges in distribution as $r\to\infty$ to $X^*(\cdot)$, where

$$X^*(t) = \frac{(W^*(t)-K\beta_e)^+}{\beta} + \frac{W^*(t)\wedge K\beta_e}{\beta_e} \qquad \text{for } t\ge 0,$$

and $W^*(\cdot)$ is the reflected Brownian motion as in Proposition 2.1.

PROOF. Since $\hat{X}^r(\cdot) = \langle 1, \hat{\mathcal{Q}}^r(\cdot)+\hat{\mathcal{Z}}^r(\cdot)\rangle$ and the mapping $\Phi:\mathbf{M}\times\mathbf{M}\to\mathbb{R}$ defined by $\Phi(\nu_1,\nu_2) = \langle 1,\nu_1+\nu_2\rangle$ for any $(\nu_1,\nu_1)\in\mathbf{M}\times\mathbf{M}$ is continuous, the result follows from Theorem 2.1 and the continuous mapping theorem. □

REMARK 2.1. In other words, $X^*(\cdot)$ is a reflected Brownian motion with drift $\frac{-\theta}{\beta}$ and variance $\frac{c_a^2+c_s^2}{\beta^2}$ when it is above $K\beta_e$ and with drift $\frac{-\theta}{\beta_e}$ and variance $\frac{c_a^2+c_s^2}{\beta_e^2}$ when it is below $K\beta_e$.

2.3. *Outline of proof.* The major step to prove our main result is to establish the following *state-space collapse* result.

THEOREM 2.2. *Assume (2.14)–(2.23). Fix $T > 0$. If (2.24) holds, then*

$$\sup_{t\in[0,T]} \mathbf{d}[(\hat{\mathcal{Q}}^r(t), \hat{\mathcal{Z}}^r(t)), \Delta_{K,\nu}\hat{W}^r(t)] \Rightarrow 0 \qquad \text{as } r \to \infty.$$

The state-space collapse result is appealing, since it rigorously shows that all performance processes can be described as a simple, deterministic function of the workload process. We now use Theorem 2.2 to prove the main result.

PROOF OF THEOREM 2.1. We have the convergence of the workload processes $\hat{W}^r(t)$ in Proposition 2.1. Since the mapping $\Delta_{K,\nu} : \mathbb{R}_+ \to \mathbf{M} \times \mathbf{M}$ is continuous, by the continuous mapping theorem

$$\Delta_{K,\nu}\hat{W}^r(\cdot) \Rightarrow \Delta_{K,\nu}W^*(\cdot) \qquad \text{as } r \to \infty.$$

The result of the theorem follows immediately from the state-space collapse result in Theorem 2.2 and the "convergence together lemma" (Theorem 4.1 in [4]). □

The proof of the state-space collapse is given in Section 5, but it requires ample preparation. Our proof is analogous to the framework developed in Bramson [6] for proving state-space collapse in multi-class networks with head-of-the-line service disciplines. The framework was later adopted for the PS queue in Gromoll [12], which the current paper closely follows. In Section 3, we establish several fundamental properties for the equilibrium behavior of the LPS fluid model introduced in [29]. Section 4 establishes precompactness of a family of shifted fluid scaled processes, which will be defined in that section. Briefly speaking, the proof of the state-space collapse is built on the "richness" of the set of fluid limits, which are obtained from the shifted fluid scaled processes. Each fluid limit is shown to be a fluid model solution in Section 5.1.

3. Convergence to equilibrium states for fluid model. We propose a fluid model, denoted by (K, λ, ν), to assist the study of the underlying stochastic processes for the LPS queue. The parameters K, λ and ν are the limiting sharing level defined in (2.19), the limiting arrival rate defined in (2.15), and the limiting job size distribution defined (2.16), respectively. According to condition (2.15), (2.16), (2.17) and (2.23), we have that the traffic intensity of the fluid model

$$\rho = \lambda\beta = 1. \tag{3.1}$$

Although ν is required to satisfy (2.18) in Theorem 2.1, in this section only, we allow the job size distribution to have atoms. The fluid analogue of the LPS queue was first proposed and studied in [29], where general properties of the fluid model were studied for all traffic intensities $\rho \in [0, \infty)$. For the purpose of this paper, we now recall the definition and some general properties for critically loaded fluid model.

Given a measure-valued process $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M}\times\mathbf{M})$, for $t \geq 0$, let

$$\bar{Q}(t) = \langle 1, \bar{\mathcal{Q}}(t)\rangle, \tag{3.2}$$

$$\bar{Z}(t) = \langle 1, \bar{\mathcal{Z}}(t)\rangle, \tag{3.3}$$

$$\bar{X}(t) = \bar{Q}(t) + \bar{Z}(t), \tag{3.4}$$

$$\bar{B}(t) = \lambda t - \bar{Q}(t). \tag{3.5}$$

These quantities are the fluid analogues of $Q(t), Z(t), B(t)$ and $X(t)$ in the stochastic model. Define the *fluid cumulative service amount* up to time t by

$$\bar{S}(t) = \int_0^t \phi_\rho(\bar{Z}(\tau))\, d\tau, \tag{3.6}$$

where

$$\phi_\rho(x) = \begin{cases} 1/x, & x > 0, \\ \infty, & x = 0, \end{cases} \tag{3.7}$$

when $\rho = 1$. It worth noting that the function ϕ_ρ take a slightly different form when $\rho \neq 1$, which is not the case in this paper. Interested readers are referred to [29] for detailed discussion. And for $0 \leq s \leq t$, denote

$$\bar{S}(s,t) = \int_s^t \phi_\rho(\bar{Z}(\tau))\, d\tau. \tag{3.8}$$

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

$$\xi = (\langle 1, \xi\rangle + \langle 1, \mu\rangle - K)^+\nu, \tag{3.9}$$

$$\langle 1, \mu\rangle = (\langle 1, \xi\rangle + \langle 1, \mu\rangle) \wedge K. \tag{3.10}$$

Roughly speaking, validity of an initial state means that the initial state is consistent with the LPS policy; initial waiting jobs have the same job size distribution as arriving jobs. Denote

$$\mathscr{I} = \{(\xi,\mu) \in \mathbf{M}\times\mathbf{M} : (\xi,\mu) \text{ satisfies (3.9) and (3.10)}\} \tag{3.11}$$

the set of all valid initial conditions.

We now introduce the following *fluid dynamic equations*, which are analogous to (2.5) and (2.6). For all $A_y = (y,\infty)$, $y \ge 0$,

$$\bar{\mathcal{Q}}(t)(A_y) = \xi(A_y) + (\bar{Q}(t) - \bar{Q}(0))\nu(A_y), \tag{3.12}$$

$$\bar{\mathcal{Z}}(t)(A_y) = \mu(A_y + \bar{S}(t)) + \int_0^t \nu(A_y + \bar{S}(s,t))\, d\bar{B}(s), \tag{3.13}$$

where $\bar{Q}(\cdot)$, $\bar{Z}(\cdot)$, $\bar{X}(\cdot)$, $\bar{B}(\cdot)$ and $\bar{S}(\cdot)$ are defined in (3.2)–(3.8). They are subject to the following constraints:

$$\bar{B}(\cdot) \text{ is nondecreasing,} \tag{3.14}$$

$$\bar{Q}(t) = (\bar{X}(t) - K)^+, \tag{3.15}$$

$$\bar{Z}(t) = (\bar{X}(t) \wedge K). \tag{3.16}$$

Because $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M}\times\mathbf{M})$, $\bar{Q}(\cdot)$, $\bar{X}(\cdot)$, $\bar{Z}(\cdot)$ and $\bar{B}(\cdot)$ are right continuous on $[0,\infty)$ and have left limits in $(0,\infty)$. Here and later, the integral $\int_0^t g(s)\, d\bar{B}(s)$ is interpreted as the Lebesgue–Stieltjes integral on the interval $[0,t]$, where by convention we set $B(0^-) = B(0)$. The above dynamic equations and constraints define the fluid model (K,λ,ν).

Definition 3.1. $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \mathbf{D}([0,\infty), \mathbf{M}\times\mathbf{M})$ is a solution to the fluid model (K,λ,ν) with a valid initial condition (ξ,μ) if it satisfies the fluid dynamic equations (3.12) and (3.13), subject to the constraints (3.14)–(3.16).

Note that (2.17) and (2.17) imply that

$$\beta < \infty, \tag{3.17}$$

and (2.18) implies

$$\nu(\{0\}) = 0, \tag{3.18}$$

$$\nu \text{ is nonlattice.} \tag{3.19}$$

It has been proved in [29] (cf. Theorem 3.1) that under the conditions (3.17) and (3.18), there exists a unique fluid model solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ for any valid initial condition (ξ,μ). Moreover, by Proposition 3.1 in [29], the fluid workload process which is defined as $\bar{W}(t) = \langle \chi, \bar{\mathcal{Q}}(t) + \bar{\mathcal{Z}}(t)\rangle$ satisfies the workload conservation property, that is,

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

Since we restrict to the critically loaded case, that is, $\rho = 1$, we have

$$\bar{W}(t) = \langle \chi, \xi + \mu\rangle \qquad \text{for all } t \geq 0. \tag{3.20}$$

The main objective of this section is to show the following long-term behavior of the critically loaded fluid model, which helps to establish the state-space collapse in Section 5.

THEOREM 3.1. *Assume (3.1) and (3.17)–(3.19). The unique solution $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ to the fluid model (K, λ, ν) with a valid initial state (ξ, μ) such that $w = \langle \chi, \xi + \mu\rangle < \infty$ satisfies*

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) \to \Delta_{K,\nu} w \qquad \textit{as } t \to \infty.$$

Moreover, for fixed constants $p, M > 0$ the convergence is uniform for all fluid solutions with initial conditions in the set

$$\mathscr{I}_M^p = \{(\xi, \mu) \in \mathscr{I} : \langle \chi, \xi + \mu\rangle < M, \langle \chi^{1+p}, \xi + \mu\rangle < M\}. \tag{3.21}$$

Section 3.1 characterizes the equilibrium states for the fluid model. Section 3.2 presents the proof of convergence (the first half of Theorem 3.1), and Section 3.3 presents the proof of uniform convergence (the second half of Theorem 3.1).

3.1. *Equilibrium states.*

DEFINITION 3.2. An element $(\xi, \mu) \in \mathscr{I}$ is called an *equilibrium state* for the fluid model (K, λ, ν) if the solution to the fluid model with initial condition (ξ, μ) satisfies

$$(\bar{\mathcal{Q}}(t), \bar{\mathcal{Z}}(t)) = (\xi, \mu) \qquad \text{for all } t \geq 0.$$

The simple intuition here is that if the fluid model solution starts with an invariant state, it will stay at this state forever. By the restarting lemma, Lemma 4.2 in [29], a fluid model solution will remain in an equilibrium state once reach it. Our first result is a characterization of an equilibrium state.

THEOREM 3.2. *An element $(\xi, \mu) \in \mathscr{I}$ is an equilibrium state if and only if*

$$(\xi, \mu) = \Delta_{K,\nu} w \qquad \textit{for some } w \in [0, \infty). \tag{3.22}$$

PROOF. Suppose $(\xi,\mu)=\Delta_{K,\nu}w$ for some $w\in[0,\infty)$, we need show that

$$(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\equiv\Delta_{k,\nu}w=\left(\frac{(w-K\beta_e)^+}{\beta}\nu,\frac{w\wedge K\beta_e}{\beta_e}\nu_e\right)$$

is the fluid model solution. If $w=0$, then by weak stability (Theorem 3.2 in [29]), $\Delta_{k,\nu}0=(\mathbf{0},\mathbf{0})$ is the fluid model solution. So let us now focus on the case where $w>0$. The fluid amount of jobs in buffer size and in service are

$$\bar{Q}(t)=\langle 1,\bar{\mathcal{Q}}(t)\rangle=\frac{(w-K\beta_e)^+}{\beta},$$
$$\bar{Z}(t)=\langle 1,\bar{\mathcal{Z}}(t)\rangle=\frac{w\wedge K\beta_e}{\beta_e}.$$

If $\bar{Z}(t)<K$, then $w<K\beta_e$ which implies that $\bar{Q}(t)=0$; if $\bar{Q}(t)>0$, then $w>K\beta_e$ which implies that $\bar{Z}(t)=K$. So condition (3.15) and (3.16) in Definition 3.1 are satisfied. Since $\bar{Q}(t)$ and $\bar{Z}(t)$ remain to be a constant, (3.14) holds trivially. This also implies that the fluid dynamic equation (3.12) is satisfied. It remains to verify the fluid dynamic equation (3.13). The fluid accumulative service amount

$$\bar{S}(t)=\frac{t}{\bar{Z}(0)}=\frac{\beta_e}{w\wedge K\beta_e}t$$

since $\bar{Z}(t)$ is a constant. The right-hand side of (3.13) becomes

$$\frac{w\wedge K\beta_e}{\beta_e}\nu_e\left(A_y+\frac{\beta_e}{w\wedge K\beta_e}t\right)+\lambda\int_0^t\nu\left(A_y+\frac{\beta_e}{w\wedge K\beta_e}(t-s)\right)ds,$$

which equals $\frac{w\wedge K\beta_e}{\beta_e}\nu_e(A_y)=\bar{\mathcal{Z}}(t)(A_y)$, for all $y\geq 0$. So (3.13) is verified. Thus, $(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))$ is the fluid model solution.

Suppose that (ξ,μ) is an equilibrium state, we need to show that (ξ,μ) takes the form (3.22). If $(\xi,\mu)=(\mathbf{0},\mathbf{0})$, then trivially $(\xi,\mu)=\Delta_{K,\nu}0$. Let us now assume that $(\xi,\mu)\neq(\mathbf{0},\mathbf{0})$. Since $(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\equiv(\xi,\mu)$ is the fluid model solution, the fluid dynamic equation (3.13) must be satisfied. That is,

$$\begin{aligned}\mu(A_y)&=\mu\left(A_y+\frac{t}{\langle 1,\mu\rangle}\right)+\lambda\int_0^t\nu\left(A_y+\frac{t-s}{\langle 1,\mu\rangle}\right)ds,\\&=\mu\left(A_y+\frac{t}{\langle 1,\mu\rangle}\right)+\frac{\langle 1,\mu\rangle}{\beta}\int_y^{y+t/\langle 1,\mu\rangle}\nu(A_{s'})\,ds'\end{aligned}$$

for all $y,t\geq 0$. This yields

$$\mu(A_y)-\mu\left(A_y+\frac{t}{\langle 1,\mu\rangle}\right)=\langle 1,\mu\rangle\left(\nu_e(A_y)-\nu_e\left(A_y+\frac{t}{\langle 1,\mu\rangle}\right)\right),$$

which implies that $\mu = \langle 1, \mu\rangle \nu_e$ due to the arbitrary of t and y. Since (ξ, μ) is a valid state, $\xi = \langle 1, \xi\rangle \nu$. Let

$$w = \langle \chi, \xi + \mu\rangle = \langle 1, \xi\rangle \beta + \langle 1, \mu\rangle \beta_e.$$

Again by validity of state (ξ, μ), $\langle 1, \xi\rangle = \frac{(w - K\beta_e)^+}{\beta}$ and $\langle 1, \mu\rangle = \frac{w \wedge K\beta_e}{\beta_e}$. So we conclude that $(\xi, \mu) = \Delta_{K,\nu} w$. $\square$

3.2. *Convergence to equilibrium states.* We now identify conditions under which the fluid model solution starting at a valid initial state (ξ, μ) will converge to an equilibrium state.

If the initial condition $(\xi, \mu) = (\mathbf{0}, \mathbf{0})$, then by weak stability (Theorem 3.2 in [29]), the fluid model solution will always be zero. So $(\mathbf{0}, \mathbf{0})$ is an equilibrium state. From now on, we focus on the case where the initial condition $(\xi, \mu) \neq (\mathbf{0}, \mathbf{0})$. By the fluid dynamic equation (3.12), $\bar{Q}(t)\beta = \langle \chi, \bar{\mathcal{Q}}(t)\rangle$. It follows from the workload conservation property (3.20) that $w = \langle \chi, \xi + \mu\rangle \equiv \bar{W}(t) \geq \langle \chi, \bar{\mathcal{Q}}(t)\rangle$ for all $t \geq 0$. So

$$\bar{Q}(t) = (\bar{X}(t) - K)^+ \leq \frac{w}{\beta} \qquad \text{for all } t \geq 0. \tag{3.23}$$

Since $\bar{W}(t) = 0$ if and only if $\bar{Z}(t) = 0$,

$$\bar{Z}(t) = (\bar{X}(t) \wedge K) > 0 \qquad \text{for all } t \geq 0. \tag{3.24}$$

So the function $\bar{S}(\cdot)$ as defined in (3.6) has an inverse on the interval $[0, \infty)$, which is denoted by $\bar{T}(\cdot)$. By the inverse function theorem,

$$\bar{T}'(v) = \bar{Z}(\bar{T}(v)) \qquad \text{for all } v \geq 0.$$

According to (3.13) in Definition 3.1, we have

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

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

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

Note that the function $\bar{Q}(\bar{T}(\cdot))$ has bounded variation since it is the difference of two nondecreasing function $\bar{B}(\bar{T}(\cdot))$ and $\lambda\bar{T}(\cdot)$. According to the integration by parts formula provided by Lemma A.1, we obtain

$$\begin{aligned}\bar{Z}(\bar{T}(u)) = \mu(A_u) + \lambda\beta \int_0^u \bar{Z}(\bar{T}(u - v))\, dF_e(v) - [1 - F(0)]\bar{Q}(\bar{T}(u)) \\ + [1 - F(u)]\bar{Q}(0) + \int_0^u \bar{Q}(\bar{T}(u - v))\, dF(v),\end{aligned} \tag{3.25}$$

where F_e is the equilibrium distribution of F which can be written as $F_e(x) = \frac{1}{\beta}\int_0^x [1-F(y)]\,dy$. It follows from (3.1) and (3.18) that $\rho = 1$ and $F(0) = 0$, so we obtain the following key relationship:

$$\begin{aligned}\bar{Q}(\bar{T}(u)) + \bar{Z}(\bar{T}(u)) = \xi(A_u) + \mu(A_u) + \int_0^u \bar{Q}(\bar{T}(u-v))\,dF(v)\\ + \int_0^u \bar{Z}(\bar{T}(u-v))\,dF_e(v)\end{aligned} \tag{3.26}$$

for all $0 \le u < \infty$. To simplify notation, denote

$$\begin{aligned}h_{\xi,\mu}(u) &= \xi(A_u) + \mu(A_u),\\ x(u) &= q(u) + z(u),\end{aligned} \tag{3.27}$$

where

$$q(u) = \bar{Q}(\bar{T}(u)), \tag{3.28}$$

$$z(u) = \bar{Z}(\bar{T}(u)). \tag{3.29}$$

By (3.15) and (3.16), the above equation can be written as

$$\begin{aligned}x(u) = h_{\xi,\mu}(u) + \int_0^u (x(u-v) - K)^+\,dF(v)\\ + \int_0^u (x(u-v) \wedge K)\,dF_e(v).\end{aligned} \tag{3.30}$$

By Lemma A.1 in [29], for any valid initial condition (ξ,μ), the above integral equation has a unique solution which is a càdlàg function. Our analysis on the limiting behavior of fluid model solutions will be mainly based on (3.30).

For a nonzero valid initial condition (ξ,μ) and an $\varepsilon \in (0,1)$, define the ε-perturbation of it by

$$(\xi_\varepsilon, \mu_\varepsilon) = \begin{cases} \left(\xi + \left(\varepsilon - \dfrac{K - \langle 1,\mu\rangle}{\langle 1,\mu\rangle}\right)^+ \mu, \left(1 + \dfrac{K - \langle 1,\mu\rangle}{\langle 1,\mu\rangle} \wedge \varepsilon\right)\mu\right), \\ \qquad\qquad\qquad\qquad \text{if } \langle 1,\mu\rangle < K, \\ (\xi + \varepsilon\mu, \mu), \qquad \text{if } \langle 1,\mu\rangle = K, \xi = \mathbf{0}, \\ (\xi + \varepsilon(\xi+\mu), \mu), \qquad \text{if } \xi \ne \mathbf{0}, \end{cases}$$

and the $-\varepsilon$-perturbation of it by

$$(\xi_{-\varepsilon}, \mu_{-\varepsilon}) = \begin{cases} (\xi, (1-\varepsilon)\mu), & \text{if } \xi = \mathbf{0}, \\ \left(\left(1 - \varepsilon\dfrac{\langle \chi, \xi+\mu\rangle}{\langle \chi,\xi\rangle}\right)^+ \xi, \right. \\ \quad \left((1-\varepsilon)\dfrac{\langle \chi, \xi+\mu\rangle}{\langle \chi,\mu\rangle} 1_{\{1 < \varepsilon\langle \chi,\xi+\mu\rangle/\langle \chi,\xi\rangle\}} \right. \\ \qquad \left.\left. + 1_{\{1 \ge \varepsilon\langle \chi,\xi+\mu\rangle/\langle \chi,\xi\rangle\}}\right)\mu\right), & \text{if } \xi \ne \mathbf{0}. \end{cases}$$

The simple idea behind this complicated looking construction is that (a) the perturbed state $(\xi_\varepsilon, \mu_\varepsilon)$ is still a valid initial condition, (b) the workload of the perturbed state satisfies

$$\langle \chi, \xi_\varepsilon + \mu_\varepsilon \rangle = (1+\varepsilon)\langle \chi, \xi + \mu \rangle \qquad \text{for all } \varepsilon \in (-1,1), \tag{3.31}$$

and (c) the function $h_{\xi_\varepsilon,\mu_\varepsilon}$ which is defined based on $(\xi_\varepsilon, \mu_\varepsilon)$ in the same as (3.27) satisfies

$$h_{\xi_{-\varepsilon},\mu_{-\varepsilon}}(u) \le h_{\xi,\mu}(u) \le h_{\xi_\varepsilon,\mu_\varepsilon}(u) \qquad \text{for all } u \in \mathbb{R}_+, \varepsilon \in (0,1). \tag{3.32}$$

The complication in the construction on the perturbation comes from the requirement (3.31), which will provide convenience in the proof of Lemmas 3.2 and 3.4. Inequality (3.32) will be used in the proof of the following lemma. Let $x^\varepsilon(\cdot)$ denote the solution to (3.30) with $h_{\xi,\mu}$ replaced by $h_{\xi_\varepsilon,\mu_\varepsilon}$. We have the following comparison.

LEMMA 3.1. *Assume (3.17), (3.18) and $(\xi,\mu) \ne (\mathbf{0},\mathbf{0})$. For all $\varepsilon \in (0,1)$,*

$$x^{-\varepsilon}(u) \le x(u) \le x^\varepsilon(u) \qquad \textit{for all } u \ge 0.$$

PROOF. Let $u^* = \inf\{u \ge 0 : x(u) > x^\varepsilon(u)\}$. To prove $x(u) \le x^\varepsilon(u)$, it is enough to show that $u^* = \infty$. Note that $x(0) = h_{\xi,\mu}(0) < h_{\xi_\varepsilon,\mu_\varepsilon}(0) = x^\varepsilon(0)$. By right continuity of $x(\cdot)$ and $x^\varepsilon(\cdot)$, $u^* > 0$. Now, suppose $u^* < \infty$. By (3.30) and (3.32) we have following bound estimation:

$$\begin{aligned} & x^\varepsilon(u^*) - x(u^*) \\ & \quad \ge \int_0^{u^*} [(x^\varepsilon(u^*-v) - K)^+ - (x(u^*-v) - K)^+]\, dF(v) \\ & \qquad + \int_0^{u^*} [(x^\varepsilon(u^*-v) \wedge K) - (x(u^*-v) \wedge K)]\, dF_e(v). \end{aligned} \tag{3.33}$$

Assumption (3.18) implies that $F(0) < 1$. So there exists $u' \in (0, u^*)$ such that

$$\int_{u'-\delta}^{u'} dF_e(v) > 0$$

for all $0 < \delta < u'$. By the definition of u^*, we have that

$$\kappa = x^\varepsilon(u^* - u') - x(u^* - u') > 0.$$

By right continuity of $x(\cdot)$ and $x^\varepsilon(\cdot)$, we can choose δ small enough such that

$$x^\varepsilon(v) - x(v) \ge \frac{\kappa}{2} \qquad \text{for all } u \in [u^* - u', u^* - u' + \delta].$$

So by (3.33), we have

$$x^{\varepsilon}(u^*) - x(u^*) \geq \frac{\kappa}{2} \int_{u'-\delta}^{u'} dF_e(v) > 0.$$

This contradicts the definition of u^*. So we must have that $u^* = \infty$. The proof for the other inequality is completely analogous. □

For the solution $x(\cdot)$ to (3.30) with initial condition (ξ, μ), define

$$x(\infty) = \frac{1}{\beta}(w - K\beta_e)^+ + \frac{1}{\beta_e}(w \wedge K\beta_e), \tag{3.34}$$

where $w = \langle \chi, \xi + \mu \rangle$. The quantity $x(\infty)$ can be interpreted as the fluid system size corresponding to the equilibrium state with workload w. We now use the above lemma and the key renewal theorem to show the following convergence. To help with the proof, we introduce the renewal function

$$U(x) = \sum_{n=0}^{\infty} F^{*n}(x),$$

associated with the distribution function F (see Section V.2 in [1] for detailed discussion).

Lemma 3.2. *Assume (3.1) and (3.17)–(3.19). The solution $x(\cdot)$ to (3.30) with initial condition $(\xi, \mu) \in \mathscr{I}$ and $\langle \chi, \xi + \mu \rangle < \infty$ satisfies*

$$x(u) \to x(\infty) \qquad \textit{as } u \to \infty.$$

Proof. We first study the case where $w = \langle \chi, \xi + \mu \rangle > K\beta_e$. Convolve both sides of (3.30) with, the renewal function $U(\cdot)$ of $F(\cdot)$, to get

$$x * U(u) = h_{\xi,\mu} * U(u) + (x - K)^+ * F * U(u) + (x \wedge K) * F_e * U(u).$$

Since $x = (x - K)^+ + (x \wedge K)$, by moving all terms containing $(x - K)^+$ to the left and all terms containing $(x \wedge K)$ to the right, we obtain

$$(x(u) - K)^+ * (1 - F) * U(u) = h_{\xi,\mu} * U(u) - (x \wedge K) * (1 - F_e) * U(u).$$

This gives

$$\begin{aligned}(x(u) - K)^+ &= h_{\xi,\mu} * U(u) - K(1 - F_e) * U(u) \\ &\quad + [K - (x \wedge K)] * (1 - F_e) * U(u).\end{aligned} \tag{3.35}$$

Both $h_{\xi,\mu}(\cdot)$ and $1 - F_e(\cdot)$ are directly Riemann integrable since they are nonincreasing and integrable functions. By the key renewal theorem, we have the convergence of the first two terms on the right-hand side of (3.35):

$$\lim_{u\to\infty} h_{\xi,\mu} * U(u) = \frac{w}{\beta},$$

$$\lim_{u\to\infty} K(1-F_e) * U(u) = \frac{K\beta_e}{\beta}.$$

Note that $\frac{w}{\beta} - \frac{K\beta_e}{\beta} > 0$ in this case, and the last term in (3.35) is always nonnegative. So there exists $u_1 > 0$ such that

$$(x(u) - K)^+ > 0 \qquad \text{for all } u \ge u_1.$$

Equivalently, this means that $K - (x(u) \wedge K) = 0$ for all $u \ge u_1$. So the last term in (3.35) is nonnegative and can be bounded above by

$$\int_{u-u_1}^{u} K\, d[(1-F_e) * U(v)] = K[(1-F_e) * U(u) - (1-F_e) * U(u-u_1)],$$

which converges to 0 by the key renewal theorem. So in this case we have $\lim_{t\to\infty} x(u) = K + \frac{(w-K\beta_e)}{\beta} = x(\infty)$.

In the case where $w < K\beta_e$, we convolve both sides of (3.30) with $U_e(\cdot)$, the renewal function of $F_e(\cdot)$ to get

$$x * U_e(u) = h_{\xi,\mu} * U_e(u) + (x-K)^+ * F * U_e(u) + (x \wedge K) * F_e * U_e(u).$$

Again, since $x = (x-K)^+ + (x\wedge K)$, by moving all terms containing $(x-K)^+$ to the right and all terms containing $(x \wedge K)$ to the left, we obtain

$$(x(u) \wedge K) = h_{\xi,\mu} * U_e(u) - (x-K)^+ * (1-F) * U_e(u). \tag{3.36}$$

By the key renewal theorem, the first term in the above converges:

$$\lim_{u\to\infty} h_{\xi,\mu} * U_e(u) = \frac{w}{\beta_e}.$$

Note that $\frac{w}{\beta_e} - K < 0$ in this case, and the last term in (3.36) is always nonpositive. So there exists $u_2 > 0$ such that

$$(x(u) \wedge K) < K \qquad \text{for all } u \ge u_2.$$

Equivalently, this means that $(x(u) - K)^+ = 0$ for all $u \ge u_2$. The last term in (3.36) is nonnegative and, according to the upper bound (3.23), can be bounded above by

$$\int_{u-u_2}^{u} \frac{w}{\beta}\, d[(1-F) * U_e(v)] = \frac{w}{\beta}[(1-F) * U_e(u) - (1-F) * U_e(u-u_2)],$$

which converges to 0 by the key renewal theorem. So in this case, we have $\lim_{t\to\infty} x(u) = \frac{w}{\beta_e} = x(\infty)$.

Now it only remains to study the case where $w = K\beta_e$. For any $\varepsilon \in (0,1)$, let $(\xi_\varepsilon, \mu_\varepsilon)$ denote the ε-perturbation of the initial condition (ξ, ε), introduced before Lemma 3.1. It follows from (3.31) that

$$w_\varepsilon = \langle \chi, \xi_\varepsilon + \mu_\varepsilon \rangle = (1+\varepsilon)\beta_e K.$$

Following from the discussion of our first case:

$$\lim_{u\to\infty} x^\varepsilon(u) = K + \frac{\varepsilon \beta_e K}{\beta}.$$

By Lemma 3.1, $x(u) \le x^\varepsilon(u)$ for all $u \ge 0$. So for all $\varepsilon > 0$ there exists u_1' such that when $u \ge u_1'$

$$x(u) \le K + \frac{\varepsilon\beta_e K}{\beta} + \varepsilon = K + \left(\frac{\beta_e K}{\beta} + 1\right)\varepsilon.$$

Similarly, we have the $-\varepsilon$-perturbation $(\xi_{-\varepsilon}, \mu_{-\varepsilon})$ for $-\varepsilon \in (-1, 0)$. It follows from (3.31) that

$$w_{-\varepsilon} = \langle \chi, \xi_{-\varepsilon} + \mu_{-\varepsilon} \rangle = (1-\varepsilon)\beta_e K.$$

Following from the discussion of our first case,

$$\lim_{u\to\infty} x^{-\varepsilon}(u) = K - \varepsilon K.$$

By Lemma 3.1, $x(u) \ge x^{-\varepsilon}(u)$ for all $u \ge 0$. So for all $\varepsilon > 0$ there exists u_2' such that when $u \ge u_2'$

$$x(u) \ge K - \varepsilon K - \varepsilon = K - (K+1)\varepsilon.$$

Summarizing this case, we have $\lim_{t\to\infty} x(u) = K = x(\infty)$. $\square$

LEMMA 3.3. *Assume (3.1) and (3.17)–(3.19). Let $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot))$ be the solution to the fluid model (K, λ, ν) with initial condition $(\xi, \mu) \in \mathscr{I}$ and $\langle \chi, \xi + \mu \rangle < \infty$. Let $w = \langle \chi, \mu \rangle$. We have as $t \to \infty$,*

$$\mathbf{d}\left[\bar{\mathcal{Q}}(t), \frac{(w - K\beta_e)^+}{\beta}\nu\right] \to 0, \tag{3.37}$$

$$\sup_{y\in(0,\infty)} \left| \bar{\mathcal{Z}}(t)(A_y) - \frac{w \wedge K\beta_e}{\beta_e}\nu_e(A_y)\right| \to 0. \tag{3.38}$$

PROOF. If $w = 0$, the result holds trivially. Now assume that $w \ne 0$. Let

$$q(\infty) = (x(\infty) - K)^+ = \frac{(w - K\beta_e)^+}{\beta},$$

$$z(\infty) = x(\infty) \wedge K = \frac{w \wedge K\beta_e}{\beta_e},$$

where $x(\infty)$ is defined in (3.34). Based on the fluid dynamic equation (3.12) and the fact that ν is a probability measure, we have

$$|\bar{\mathcal{Q}}(t)(A) - q(\infty)\nu(A)| = |[\bar{Q}(t) - q(\infty)]\nu(A)| \le |\bar{Q}(t) - q(\infty)|,$$

for all Borel set A. This implies that

$$\bar{\mathcal{Q}}(t)(A) \le q(\infty)\nu(A) + |\bar{Q}(t) - q(\infty)| \le q(\infty)\nu(A^{|\bar{Q}(t)-q(\infty)|}) + |\bar{Q}(t) - q(\infty)|,$$

where A^κ denote the κ-enlargement as introduced in Section 1.1. Similarly, we have

$$q(\infty)\nu(A) \le \bar{\mathcal{Q}}(t)(A^{|\bar{Q}(t)-q(\infty)|}) + |\bar{Q}(t) - q(\infty)|.$$

By the change of variable $u = \bar{S}(t)$ $[t = \bar{T}(u)]$ and the definition of the Prohorov metric,

$$\mathbf{d}[\bar{\mathcal{Q}}(t), q(\infty)\nu] \le |q(u) - q(\infty)|.$$

By Lemma 3.2, there exists $u_1 > 0$ such that when $u > u_1$ we have $|q(u) - q(\infty)| \le \varepsilon$. So for all $\varepsilon > 0$, there exists $t_1 = Ku_1 \ge \bar{T}(u_1)$ such that

$$\mathbf{d}[\bar{\mathcal{Q}}(t), q(\infty)\nu] \le \varepsilon \qquad \text{for all } t > t_1. \tag{3.39}$$

It remains to study the limiting behavior of $\bar{\mathcal{Z}}(\cdot)$. Perform the change of variable $u = \bar{S}(t)$ $(t = \bar{T}(u))$ to the fluid dynamic equation (3.13), we get

$$\bar{\mathcal{Z}}(\bar{T}(u))(A_y) = \mu(A_y + u) + \int_0^u \nu(A_y + u - v)\, d[\lambda \bar{T}(v) - q(v)].$$

Due to the fact that $\rho = 1$, we have $\frac{1}{\beta} = \lambda$. Thus, $z(\infty)\nu_e(A_y) = \lambda \int_0^u \nu(A_y + u - v) z(\infty)\, dv + z(\infty)\nu_e(A_y + u)$. Since $d\bar{T}(v) = z(v)\, dv$, we then have the following bound estimation:

$$\begin{aligned} &|\bar{\mathcal{Z}}(\bar{T}(u))(A_y) - z(\infty)\nu_e(A_y)| \\ &\quad \le \mu(A_y + u) + \left| \int_0^u \nu(A_y + u - v)\, dq(v) \right| \\ &\qquad + \left| \lambda \int_0^u \nu(A_y + u - v)[z(\infty) - z(v)]\, dv \right| \\ &\qquad + z(\infty)\nu_e(A_y + u). \end{aligned} \tag{3.40}$$

It is clear that the first and the last terms on the right-hand side of (3.40) vanishes as $u \to \infty$. Recall that F is the distribution function corresponding to the measure ν. By integration by parts (see Lemma A.1), the second term on the right-hand side of (3.40) can be written as

$$\left| [1 - F(y)]q(u) - [1 - F(y + u)]q(0) + \int_0^u q(v)\, dF(y + u - v) \right|,$$

which is less than or equal to

$$|[1-F(y)]q(u)-[F(y+u)-F(y)]q(\infty)|+|[1-F(y+u)]q(0)| \\ +\int_0^u |q(u-v)-q(\infty)|\,dF(y+v).$$

By convergence of $x(\cdot)$, for all $\varepsilon>0$ there exists a u_1 such that $|x(v)-x(\infty)|<\varepsilon$ if $v\geq u_1$. For all $\varepsilon>0$, we can choose $u_2>0$ such that $1-F(u_2)<\varepsilon$. When $u>u_1+u_2$, the above inequality can be further bounded by

$$\begin{aligned}
&[1+q(0)+q(\infty)]\varepsilon+\int_0^{u-u_1}|q(u-v)-q(\infty)|\,dF(y+v)\\
&\quad+\int_{u-u_1}^{u}|q(u-v)-q(\infty)|\,dF(y+v)\\
&\qquad\leq[1+q(0)+q(\infty)]\varepsilon+[F(u-u_1+y)-F(y)]\varepsilon\\
&\qquad\quad+\sup_{v\geq 0}|q(v)-q(\infty)|[F(u)-F(u-u_1)]\\
&\qquad\leq\left(2+\frac{2w}{\beta}\right)\varepsilon,
\end{aligned}$$

where the last inequality is due to (3.23). When $u>u_1+u_2$, the third term in (3.40) can be written as

$$\left|\lambda\int_0^{u_1}\nu(A_y+u-v)[z(\infty)-z(v)]\,dv+\lambda\int_{u_1}^{u}\nu(A_y+u-v)[z(\infty)-z(v)]\,dv\right|,$$

which is bounded above by

$$\begin{aligned}
&\lambda\sup_{0\leq v\leq u_1}|z(v)-z(\infty)|[1-F(u-u_1)]+\varepsilon\lambda\int_{u_1}^{u}[1-F(v)]\,dv\\
&\qquad\leq\lambda K\varepsilon+\lambda\beta\varepsilon,
\end{aligned}$$

where the last inequality is due to the bound $z(u)\leq K$ for all $u\geq 0$. So for all $\varepsilon>0$ there exists a $t_2=K(u_1+u_2)\geq\bar{T}(u_1+u_2)$ such that

$$\sup_{y\in(0,\infty)}|\bar{\mathcal{Z}}(t)(A_y)-z(\infty)\nu_e(A_y)|<\varepsilon\qquad\text{for all } t\geq t_2. \tag{3.41}$$

□

Proof of Theorem 3.1, part I. Note that $\langle\chi,\frac{w\wedge K\beta_e}{\beta_e}\nu_e\rangle\leq K\beta_e<\infty$. By the workload conversing property, $\langle\chi,\bar{\mathcal{Z}}(\cdot)\rangle\leq w<\infty$. According to Lemma C.1, (3.38) in Lemma 3.3 implies that

$$\mathbf{d}\left[\bar{\mathcal{Z}}(t),\frac{w\wedge K\beta_e}{\beta_e}\nu_e\right]\to 0\qquad\text{as } t\to\infty.$$

This and (3.37) implies the convergence result in Theorem 3.2. □

3.3. *Uniform convergence to equilibrium states.* The convergence in the previous subsection depends on the initial condition (ξ,μ). We now show that the convergence is uniform for all initial conditions in the set $\mathscr{I}_M^p$ defined in Theorem 3.1.

To emphasize the dependency on the initial condition, we use $\Upsilon(\mathscr{I}_M^p)$ to denote the set of solutions to equation (3.30) with input function $h_{\xi,\mu}$ induced by initial condition $(\xi,\mu)\in\mathscr{I}_M^p$, and $\Xi(\mathscr{I}_M^p)$ to denote the set of solutions to the fluid model (K,λ,ν) with initial condition $(\xi,\mu)\in\mathscr{I}_M^p$.

LEMMA 3.4. *Assume (3.1) and (3.17)–(3.19). For each $\varepsilon>0$ there exists an $l^*>0$ such that when $u\ge l^*$,*

$$\sup_{x(\cdot)\in\Upsilon(\mathscr{I}_M^p)}|x(u)-x(\infty)|<\varepsilon.$$

PROOF. To prove this lemma, we need to adjust the proof of Lemma 3.2 with the assistance of Lemma B.1.

Let $\mathscr{H}_M=\{h_{\xi,\mu}:(\xi,\mu)\in\mathscr{I}_M^p\}$. By the definition of the set $\mathscr{I}_M^p$ in Theorem 3.1, $\mathscr{H}_M$ is the set of nonincreasing functions satisfying condition (B.1) and (B.2) in Lemma B.1. For any $\varepsilon>0$, divide the set $\mathscr{I}_M^p$ into three parts,

$$\mathscr{I}_M^p=\mathscr{I}_\varepsilon^+\cup\mathscr{I}_\varepsilon^0\cup\mathscr{I}_\varepsilon^-,$$

where

$$\mathscr{I}_\varepsilon^+=\{(\xi,\mu)\in\mathscr{I}_M^p:\langle\chi,\xi+\mu\rangle\ge K\beta_e(1+\varepsilon)\},$$
$$\mathscr{I}_\varepsilon^-=\{(\xi,\mu)\in\mathscr{I}_M^p:\langle\chi,\xi+\mu\rangle\le K\beta_e(1-\varepsilon)\},$$

and $\mathscr{I}_\varepsilon^0=\mathscr{I}_M^p\setminus(\mathscr{I}_\varepsilon^+\cup\mathscr{I}_\varepsilon^-)$.

We first focus on the set $\mathscr{I}_\varepsilon^+$. By doing the same algebra as in the proof of Lemma 3.2, we see that (3.35) holds for any $(\xi,\mu)\in\mathscr{I}_\varepsilon^+$. By Lemma B.1 and the key renewal theorem, there exists a u_1^* such that

$$\sup_{(\xi,\mu)\in\mathscr{I}_\varepsilon^+}\left|h_{\xi,\mu}*U(u)-\frac{\langle\chi,\xi+\mu\rangle}{\beta}\right|<\frac{K\beta_e}{4\beta}\varepsilon,$$
$$\left|K(1-F_e)*U(u)-K\frac{\beta_e}{\beta}\right|<\frac{K\beta_e}{4\beta}\varepsilon,$$

for all $u\ge u_1^*$. So for the first two terms on the right-hand side of (3.35), we have

$$h_{\xi,\mu}*U(u)-K(1-F_e)*U(u)\ge\frac{K\beta_e(1+\varepsilon)}{\beta}-K\frac{\beta_e}{\beta}-\frac{K\beta_e}{2\beta}\varepsilon>0,$$

for all $(\xi,\mu)\in\mathscr{I}_\varepsilon^+$ and $u>u_1^*$. Note that the last term in (3.35) is always nonnegative. So when $u\ge u_1^*$ we have $(x(u)-K)^+>0$ [or equivalently $K-$

$(x(u) \wedge K) = 0]$, for all $(\xi, \mu) \in \mathscr{I}_\varepsilon^+$. So the last term on the right-hand side of (3.35) is nonnegative and can be bounded above by

$$\int_{u-u_1^*}^{u} K\, d[(1-F_e) * U(v)] = K[(1-F_e) * U(u) - (1-F_e) * U(u-u_1^*)],$$

which converges to 0 as $u \to \infty$ by the key renewal theorem. So there exists a $u_1' > 0$ such that when $u > u_1'$, the absolute value of third term on the right-hand side of (3.35) is bounded by $\frac{K\beta_e}{2\beta}\varepsilon$. Let $l_1^* = \max(u_1', u_1^*)$. By (3.35) and summarizing the above, we obtain

$$\sup_{(\xi,\mu)\in\mathscr{I}_\varepsilon^+} |x(u) - x(\infty)| < \frac{K\beta_e}{\beta}\varepsilon \qquad \text{for all } u > l_1^*.$$

Next, we consider the set $\mathscr{I}_\varepsilon^-$. By doing the same algebra as in the proof of Lemma 3.2, we see that (3.36) holds for any $(\xi, \mu) \in \mathscr{I}_\varepsilon^-$. By Lemma B.1, there exists a u_2^* such that

$$\sup_{(\xi,\mu)\in\mathscr{I}_\varepsilon^-} \left| h_{\xi,\mu} * U_e(u) - \frac{\langle \chi, \xi + \mu\rangle}{\beta_e} \right| < \frac{K}{2}\varepsilon$$

for all $u > u_2^*$. So we have

$$h_{\xi,\mu} * U_e(u) \le \frac{K\beta_e(1-\varepsilon)}{\beta_e} + \frac{K}{2}\varepsilon < K,$$

for all $(\xi, \mu) \in \mathscr{I}_\varepsilon^-$ and $u > u_2^*$. Note that the last term in (3.36) is always nonpositive. So when $u \ge u_2^*$ we have $x(u) < K$ [or equivalently, $(x(u) - K)^+ = 0$] for all $(\xi, \mu) \in \mathscr{I}_\varepsilon^-$. So by (3.23), the absolute value of the last term on the right-hand side of (3.36) can be bounded by

$$\int_{u-u_2^*}^{u} \frac{w}{\beta}\, d[(1-F) * U_e(v)] = \frac{w}{\beta}[(1-F) * U_e(u) - (1-F) * U_e(u-u_2^*)],$$

which converges to 0 as $u \to \infty$ by the key renewal theorem. So there exists a $u_2' > 0$ such that when $u > u_2'$, the last term on the right-hand side of (3.36) is bounded by $\frac{K}{2}\varepsilon$. Let $l_2^* = \max(u_2', u_2^*)$. By (3.36) and summarizing the above,

$$\sup_{(\xi,\mu)\in\mathscr{I}_\varepsilon^-} |x(u) - x(\infty)| < K\varepsilon \qquad \text{for all } u > l_2^*.$$

It only remains to deal with the set $\mathscr{I}_\varepsilon^0$. We can restrict $\varepsilon < 1/3$, since we are only interested in small ones. According to (3.31), for any $(\xi, \mu) \in \mathscr{I}_\varepsilon^0$, we have $h_{\xi_{3\varepsilon},\mu_{3\varepsilon}} \in \mathscr{I}_\varepsilon^+$ and $h_{\xi_{-3\varepsilon},\mu_{-3\varepsilon}} \in \mathscr{I}_\varepsilon^-$. Denote $x^+(\cdot)$ and $x^-(\cdot)$ the solutions to (3.30) corresponding to $h_{\xi_{3\varepsilon},\mu_{3\varepsilon}}$ and $h_{\xi_{-3\varepsilon},\mu_{-3\varepsilon}}$, respectively. By Lemma 3.1,

$$x^-(u) < x(u) < x^+(u) \qquad \text{for all } u \ge 0.$$

Note that in this case, the workload $w \le (1+\varepsilon)\beta_e K \le 2\beta_e K$. By (3.34), we have that

$$x^+(\infty) \le x(\infty) + \max\left(\frac{1}{\beta}, \frac{1}{\beta_e}\right) 3\varepsilon w \le x(\infty) + \max\left(\frac{1}{\beta}, \frac{1}{\beta_e}\right) 6\beta_e K\varepsilon,$$

$$x^-(\infty) \ge x(\infty) - \max\left(\frac{1}{\beta}, \frac{1}{\beta_e}\right) 3\varepsilon w \ge x(\infty) - \max\left(\frac{1}{\beta}, \frac{1}{\beta_e}\right) 6\beta_e K\varepsilon.$$

According to the above two cases, when $u > l^* = \max(l_1^*, l_2^*)$,

$$x(u) \le x^+(\infty) + \frac{K\beta_e}{\beta}\varepsilon \le x(\infty) + \max\left(\frac{1}{\beta}, \frac{1}{\beta_e}\right) 6\beta_e K\varepsilon + \frac{K\beta_e}{\beta}\varepsilon,$$

$$x(u) \ge x^-(\infty) - K\varepsilon \ge x(\infty) - \max\left(\frac{1}{\beta}, \frac{1}{\beta_e}\right) 6\beta_e K\varepsilon - K\varepsilon,$$

for all $(\xi, \mu) \in \mathscr{I}_\varepsilon^0$. This means that

$$\sup_{(\xi,\mu)\in\mathscr{I}_\varepsilon^0} |x(u) - x(\infty)| < C\varepsilon \qquad \text{for all } u > l^*,$$

where $C = \max(\frac{1}{\beta}, \frac{1}{\beta_e}) 6\beta_e K + \frac{K\beta_e}{\beta} + K$. □

LEMMA 3.5. *Assume (3.1) and (3.17)–(3.19). For all $\varepsilon > 0$ there exists an $L^* > 0$ such that when $t \ge L^*$,*

$$\sup_{(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\in\Xi(\mathscr{I}_M^p)} \mathbf{d}\left[\bar{\mathcal{Q}}(t), \frac{(w - K\beta_e)^+}{\beta}\nu\right] < \varepsilon, \tag{3.42}$$

$$\sup_{(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\in\Xi(\mathscr{I}_M^p)} \sup_{y\in(0,\infty)} \left|\bar{\mathcal{Z}}(t)(A_y) - \frac{w \wedge K\beta_e}{\beta_e}\nu_e(A_y)\right| < \varepsilon. \tag{3.43}$$

PROOF. The proof of this corollary is almost the same as the proof of Lemma 3.3. Just note that by Lemma 3.4, the t_1 in (3.39) and the t_2 in (3.41) are good for all $(\xi, \mu) \in \mathscr{I}_M^p$. With $L^* = \max(t_1, t_2)$, the result of this lemma immediately follows. □

PROOF OF THEOREM 3.1, PART II. Now we use Lemma 3.5 to show the uniform convergence result. Note that $\langle \chi, \frac{w\wedge K\beta_e}{\beta_e}\nu_e\rangle \le K\beta_e < \infty$. By the definition of $\mathscr{I}_M^p$, for any $(\bar{\mathcal{Q}}(\cdot), \bar{\mathcal{Z}}(\cdot)) \in \Xi(\mathscr{I}_M^p)$, $\langle \chi, \bar{\mathcal{Z}}(\cdot)\rangle < M < \infty$. According to Lemma C.1, (3.43) in Lemma 3.5 implies that for all $\varepsilon > 0$ there exists an L_1^* such that when $t \ge L_1^*$,

$$\sup_{(\bar{\mathcal{Q}}(\cdot),\bar{\mathcal{Z}}(\cdot))\in\Xi(\mathscr{I}_M^p)} \mathbf{d}\left[\bar{\mathcal{Z}}(t), \frac{w \wedge K\beta_e}{\beta_e}\nu_e\right] < \varepsilon.$$

The uniform convergence follows from the above and (3.42). □

4. Shifted fluid scaling and precompactness. The objective of this section is to show the precompactness property, Theorem 4.1 at the end of this section, for the sequence of shifted fluid scaled processes, which is defined in the following section.

4.1. *Shifted fluid scaling.* Much of our understanding of the diffusion scaled process will be derived from results about the *shifted fluid scaled process*, which is defined by

$$\bar{\mathcal{Q}}^{r,m}(t) = \frac{1}{r}\mathcal{Q}^r(rm+rt), \qquad \bar{\mathcal{Z}}^{r,m}(t) = \frac{1}{r}\mathcal{Z}^r(rm+rt), \tag{4.1}$$

for all $m \in \mathbb{N}$ and $t \geq 0$. To see the relationship between these two scalings, consider the diffusion scaled process on the interval $[0, T]$, which corresponds to the interval $[0, r^2T]$ for the unscaled process. Fix a constant $L > 1$, the interval will be covered by the $\lfloor rT \rfloor + 1$ overlapping intervals

$$[rm, rm + rL], \qquad m = 0, 1, \ldots, \lfloor rT \rfloor.$$

For each $t \in [0, T]$, there exists an $m \in \{0, \ldots, \lfloor rT \rfloor\}$ and $s \in [0, L]$ (which may not be unique) such that $r^2 t = rm + rs$. Thus,

$$\hat{\mathcal{Q}}^r(t) = \bar{\mathcal{Q}}^{r,m}(s), \qquad \hat{\mathcal{Z}}^r(t) = \bar{\mathcal{Z}}^{r,m}(s). \tag{4.2}$$

This will serve as a key relationship between fluid and diffusion scaled processes.

We are also interested in shifted fluid scaled versions of other processes, like the workload and system size processes. Note that $Q^r(\cdot)$, $Z^r(\cdot)$, $X^r(\cdot)$, $W^r(\cdot)$ and $S^r(\cdot, \cdot)$ are actually functions of $(\mathcal{Q}^r(\cdot), \mathcal{Z}^r(\cdot))$, so the scaling for these quantities is defined as the functions of the corresponding scaling for $(\mathcal{Q}^r(\cdot), \mathcal{Z}^r(\cdot))$, that is,

$$\bar{Q}^{r,m}(t) = \langle 1, \bar{\mathcal{Q}}^{r,m}(t)\rangle = \frac{1}{r}Q^r(rm+rt), \tag{4.3}$$

$$\bar{Z}^{r,m}(t) = \langle 1, \bar{\mathcal{Z}}^{r,m}(t)\rangle = \frac{1}{r}Z^r(rm+rt), \tag{4.4}$$

$$\bar{X}^{r,m}(t) = \langle 1, \bar{\mathcal{Z}}^{r,m}(t) + \bar{\mathcal{Z}}^{r,m}(t)\rangle = \frac{1}{r}X^r(rm+rt), \tag{4.5}$$

$$\bar{W}^{r,m}(t) = \langle \chi, \bar{\mathcal{Q}}^{r,m}(t) + \bar{\mathcal{Z}}^{r,m}(t)\rangle = \frac{1}{r}W^r(rm+rt), \tag{4.6}$$

for all $0 \leq s \leq t$. We define the shifted fluid scaling for the arrival process as

$$\bar{E}^{r,m}(t) = \frac{1}{r}E^r(rm+rt), \tag{4.7}$$

for all $t \geq 0$. By (2.3), the shifted fluid scaling for $B^r(\cdot)$ is

$$\bar{B}^{r,m}(t) = \bar{E}^{r,m}(t) - \bar{Q}^{r,m}(t), \tag{4.8}$$

for all $t \geq 0$. To shorten the notation, for all $0 \leq s \leq t$, denote

$$\bar{E}^{r,m}(s,t) = \bar{E}^{r,m}(t) - \bar{E}^{r,m}(s), \qquad \bar{B}^{r,m}(s,t) = \bar{B}^{r,m}(t) - \bar{B}^{r,m}(s). \tag{4.9}$$

A shifted fluid scaled version of the stochastic dynamic equations (2.5) and (2.6) can be written as, for $0 \leq s \leq t$,

$$\begin{aligned} \bar{\mathcal{Q}}^{r,m}(t)(A) = \bar{\mathcal{Q}}^{r,m}(s)(A) &+ \frac{1}{r} \sum_{i=r\bar{E}^{r,m}(s)+1}^{r\bar{E}^{r,m}(t)} \delta_{v_i^r}(A) \\ &- \frac{1}{r} \sum_{i=r\bar{B}^{r,m}(s)+1}^{r\bar{B}^{r,m}(t)} \delta_{v_i^r}(A), \qquad A \subseteq [0,\infty), \end{aligned} \tag{4.10}$$

$$\begin{aligned} \bar{\mathcal{Z}}^{r,m}(t)(A) = \bar{\mathcal{Z}}^{r,m}(s)(A + S^r(rm+rs, rm+rt)) \\ + \frac{1}{r} \sum_{i=r\bar{B}^{r,m}(s)+1}^{r\bar{B}^{r,m}(t)} \delta_{v_i^r}(A + S^r(\tau_i^r, rm+rt)), \\ A \subseteq (0,\infty). \end{aligned} \tag{4.11}$$

Please note that $\bar{\mathcal{Z}}^{r,m}(t)(\{0\}) = 0$ for all $t \geq 0$ according to our definition. The dynamics of the system is determined by the above equations. Equation (4.10) 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 who left buffer enter service; the service process has been taken care of by shifting the set A by the cumulative service amount $S^r(\tau_i, rm+rt)$ that the ith job receives. This corresponds to the second term on the right-hand side of (4.11). This plus the status at time s shifted by accumulative service amount $S^r(rm+rs, rm+rt)$ is equal to the status of the server at time t, as indicated in (4.11).

4.2. *Preliminary estimates.* We first establish some bounds which will be useful for later discussion. The following lemma gives some bound on the arrival processes.

LEMMA 4.1. *Assume (2.14) and (2.15). Fix $T > 0$ and $L > 1$. For all ε, $\varepsilon' > 0$, there exists an r_0 such that whenever $r \geq r_0$,*

$$\mathbb{P}^r\Big(\max_{m \leq \lfloor rT \rfloor} \sup_{s,t \in [0,L]} |E^{r,m}(s,t) - \lambda(t-s)| > \varepsilon'\Big) < \varepsilon. \tag{4.12}$$

PROOF. Let $t' = \frac{m+t}{r}$ and $s' = \frac{m+s}{r}$. Note that $\max_{m \leq \lfloor rT \rfloor} \sup_{t \in [0,L]} \frac{m+t}{r} < T+1$ for all large r, and $0 \leq s,t \leq L$ is the same as $|t'-s'| \leq L/r$. For any

$\delta > 0$, there exists an r_0' such that $L/r < \delta$ for all $r \geq r_0$, so the left-hand side of (4.12) can be bounded above by

$$\text{(4.13)}\quad \mathbb{P}^r\left(\sup_{s',t'\in[0,T+1],|s'-t'|<\delta}\left|\frac{1}{r}E^r(r^2t') - \lambda rt' - \left(\frac{1}{r}E^r(r^2s') - \lambda rs'\right)\right| > \varepsilon'\right)$$

for all $r \geq r_0'$. By the assumptions (2.14) and (2.15) on the arrival process, $\{\frac{1}{r}E(r^2\cdot) - \lambda r\cdot\}$ converges in distribution to the Brownian motion $E^*(\cdot)$. Since a Brownian motion is almost surely continuous, we conclude that (4.13) converges to zero as $\delta \to 0$. Then the inequality (4.12) follows immediately. □

Here is a remark that will facilitate some arguments later on. The ε' and ε in (4.12) can be replaced by $\varepsilon_E(r)$, which is a function of r that vanishes at infinity. Here is the proof. For each index r let

$$H_r = \{\delta > 0 : \text{(4.12) is true for } \varepsilon' = \varepsilon = \delta\}.$$

Clearly, H_r is not empty since $2 \in H_r$. Let $\varepsilon_E(r) = \inf H_r$ for each $r \geq 0$. Assume that $\varepsilon_E(r)$ does not vanish at infinity. There exists a $\delta > 0$ and a subsequence $\{r_n\}_{n=1}^\infty$ which increases to infinity such that

$$\text{(4.14)}\qquad \varepsilon_E(r_n) > \delta \qquad \text{for all } n \geq 0.$$

However, by Lemma 4.1, for $\varepsilon' = \varepsilon = \delta/2$ there exists an r_δ such that when $r_n \geq r_\delta$, (4.12) must hold. This contradicts (4.14). Based on this, we construct

$$\text{(4.15)}\qquad \Omega_E^r = \Big\{\max_{m\leq\lfloor rT\rfloor}\sup_{s,t\in[0,L]}|E^{r,m}(s,t) - \lambda(t-s)| \leq \varepsilon_E(r)\Big\}.$$

According to Lemma 4.1, we have that

$$\text{(4.16)}\qquad \lim_{r\to\infty}\mathbb{P}^r(\Omega_E^r) = 1.$$

Recall the Glivenko–Cantelli estimate in Lemma D.1. By the same argument as in the above, for fixed constant M_1, L_1, there exists a function $\varepsilon_{\mathrm{GC}}(\cdot)$, which vanishes at infinity, such that the probability inequality in Lemma D.1 holds with ε and ε' replaced by this function. In other words, if we denote

$$\text{(4.17)}\qquad \Omega_{\mathrm{GC}}^r(M_1, L_1) = \Big\{\max_{-rM_1<n<r^2M_1}\sup_{l\in[0,L_1]}\sup_{f\in\mathscr{V}}|\langle f, \bar{\eta}^r(n,l)\rangle - l\langle f, \nu^r\rangle| \leq \varepsilon_{\mathrm{GC}}(r)\Big\},$$

where $\bar{\eta}^r(n,l)$ is defined in (D.1) and $\mathscr{V}$ is defined in (D.5), then for any fixed constant M_1, L_1,

$$\text{(4.18)}\qquad \lim_{r\to\infty}\mathbb{P}^r(\Omega_{\mathrm{GC}}^r(M_1, L_1)) = 1.$$

Now, we use the above result and Proposition 2.1 to obtain a bound on the queue length processes.

LEMMA 4.2. *Assume (2.14)–(2.17), (2.20), (2.21) and (2.23). Fix $T>0$ and $L>1$. For all $\eta>0$, there exists a constant $M>0$ such that*

$$\liminf_{r\to\infty}\mathbb{P}^r\Big(\max_{m\le\lfloor rT\rfloor}\sup_{t\in[0,L]}\bar{Q}^{r,m}(t)\le M\Big)\ge 1-\eta.$$

PROOF. Since $\frac{\lfloor rT\rfloor+L}{r}<T+1$ for all large enough r, it is enough to prove the following inequality:

$$\liminf_{r\to\infty}\mathbb{P}^r\Big(\sup_{t\in[0,T+1]}\hat{Q}^r(t)\le M\Big)\ge 1-\eta.$$

Suppose this is not true, then there exists an $\eta>0$ such that for any M,

$$\liminf_{r\to\infty}\mathbb{P}^r\Big(\sup_{t\in[0,T+1]}\hat{Q}^r(t)>M\Big)>\eta.$$

Denote the event in the above probability by Ω_1^r. By the stochastic dynamic equation (2.5), we have

$$\frac{1}{r}\mathcal{Q}(r^2t)(A)=\frac{1}{r}\sum_{i=B^r(r^2t)+1}^{E^r(r^2t)}\delta_{v_i^r}(A).$$

Since ν is a probability measure on $\mathbb{R}_+$, there exists an $a>0$ such that $\nu(a,\infty)>0$. We have the following inequality from the dynamic equation (4.10):

$$\frac{1}{r}W^r(r^2t)>a\frac{1}{r}\mathcal{Q}^r(r^2t)(a,\infty)\ge\frac{a}{r}\sum_{i=B^r(r^2t)+1}^{E^r(r^2t)}\delta_{v_i^r}(a,\infty).\tag{4.19}$$

For any r, on the event Ω_1^r there exists a $t_1\in[0,T+1]$ (random and depending on r) such that

$$\frac{1}{r}Q^r(r^2t_1)>M.$$

By (4.15), on the event Ω_E^r,

$$\sup_{t\in[0,T+1]}E^r(r^2t)\le 2\lambda r^2(T+1),$$

for all large enough r. Let $M_1=\max(M,2\lambda T)$ and $L_1=M$. By (4.17) and (4.19), on the event $\Omega_{\mathrm{GC}}^r(M_1,L_1)\cap\Omega_E^r\cap\Omega_1^r$,

$$\hat{W}^r(t_1)>aM\nu^r(a,\infty)-1>aM\nu(a,\infty)-2,$$

for all large r. By (4.16) and (4.18), we have for each $M > 0$,

$$\liminf_{r\to\infty} \mathbb{P}^r\Big(\sup_{t\in[0,T+1]} \hat{W}^r(t) > aM\nu(a,\infty) - 2\Big) > \eta.$$

This contradicts the result in Proposition 2.1. $\square$

The following lemma gives a bound on the $(1+p)$th moment of the measure-valued process, where p is the same as in conditions (2.17) and (2.21).

LEMMA 4.3. *Assume (2.14)–(2.17), (2.19)–(2.21) and (2.23). Fix $T > 0$ and $L > 1$. For each $\eta > 0$, there exists a constant $M > 0$ such that*

$$\liminf_{r\to\infty} \mathbb{P}^r\Big(\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^{r,m}(t) + \bar{\mathcal{Z}}^{r,m}(t)\rangle \le M\Big) \ge 1-\eta.$$

PROOF. By condition (2.21),

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

Denote the event in the above by Ω_0^r. By Lemma 4.2, for any $\eta > 0$, there exists a constant $M' > 0$ such that

$$\liminf_{r\to\infty} \mathbb{P}^r\left(\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \frac{1}{r} Q^r(rm+rt) \le M'\right) \ge 1-\eta/2.$$

Denote the event in the above by $\Omega_1^r(M)$. Fix $M_1 = \max(M', \lambda(T+1))$ and $L_1 = \lambda(L+1) + 2M'$. By Lemma D.1,

$$\lim_{r\to\infty} \mathbb{P}^r(\Omega_{\mathrm{GC}}^r(M_1, L_1)) = 1.$$

To prove the lemma, it suffices to show that there exists an $M > 0$ such that on the event $\Omega_0^r \cap \Omega_1^r(M') \cap \Omega_{\mathrm{GC}}^r(M_1, L_1) \cap \Omega_E^r$,

$$\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt) + \frac{1}{r}\mathcal{Z}^r(rm+rt)\right\rangle \le M,$$

for all large r. In the remainder of the proof, all random quantities of the rth system is evaluated at a sample path in the event $\Omega_0^r \cap \Omega_1^r(M') \cap \Omega_{\mathrm{GC}}^r(M_1, L_1) \cap \Omega_E^r$.

We first find a bound for $\max_{m\le\lfloor rT\rfloor} \sup_{t\in[0,L]} \langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt)\rangle$. By the dynamic equation (2.5), we have that for all $m \le \lfloor rT\rfloor$ and $t \in [0, L]$,

$$\left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt)\right\rangle = \left\langle \chi^{1+p}, \frac{1}{r}\sum_{B^r(rm+rt)}^{E^r(rm+rt)} \delta_{v_i^r}\right\rangle.$$

By (4.15) and the definition of $\Omega_1^r(M')$, we have

$$\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} E^r(rm + rt) < \lambda r^2 (T+1) \le r^2 M_1, \tag{4.20}$$

$$\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} Q^r(rm + rt) < rM' \le rL_1 \tag{4.21}$$

for all large enough r. So

$$\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} \left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Q}^r(rm+rt) \right\rangle \le \sup_{-rM_1 < n < r^2 M_1} \langle \chi^{1+p}, \bar{\eta}^r(n, L_1) \rangle.$$

By (D.5) in the remark after Lemma D.1, the function $\chi^{1+p} \in \mathscr{V}$, which appears in the definition of $\Omega_{\mathrm{GC}}^r(M_1, L_1)$. So by (4.17)

$$\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^{r,m}(t) \rangle \le L_1 \langle \chi^{1+p}, \nu^r \rangle + 1/2.$$

It then follows from condition (2.17) and Theorem 25.12 in [3], we have $\langle \chi^{1+p}, \nu^r \rangle \to \langle \chi^{1+p}, \nu \rangle$ as $r \to \infty$. Thus,

$$\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^{r,m}(t) \rangle \le L_1 \langle \chi^{1+p}, \nu \rangle + 1,$$

for all large r.

We now look for a bound for $\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0,L]} \langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(rm+rt) \rangle$. It follows from the dynamic equation (4.11) that for any $m \le \lfloor rT \rfloor$, $t \in [0,L]$ and Borel set $A \subset \mathbb{R}_+$,

$$\begin{aligned}
&\frac{1}{r}\mathcal{Z}^r(rm+rt)(A) \\
&\quad = \frac{1}{r}\mathcal{Z}^r(rm+rt)(A \cap (0,\infty)) + \frac{1}{r}\mathcal{Z}^r(rm+rt)(\{0\}) \\
&\quad = \frac{1}{r}\mathcal{Z}^r(0)(A \cap (0,\infty) + S^r(0, rm+rt)) \\
&\qquad + \sum_{j=0}^{m-1} \frac{1}{r} \sum_{i=B^r(r(m-j-1))+1}^{B^r(r(m-j))} \delta_{v_i^r}(A \cap (0,\infty) + S^r(\tau_i^r, rm+rt)) \\
&\qquad + \frac{1}{r} \sum_{i=\bar{B}^r(rm)+1}^{B^r(rm+rt)} \delta_{v_i^r}(A \cap (0,\infty) + S^r(\tau_i^r, rm+rt)).
\end{aligned}$$

Given $0 \le j \le m-1$, for those i's with $B^r(r(m-j-1)) < i \le B^r(r(m-j))$ we have

$$\tau_i^r \in [r(m-j-1), r(m-j)].$$

Let's first consider the case where $Z^r(s) > 0$ for all $s \in [0, rm+rt]$. In this case, by (2.2), the cumulative service amount

$$S^r(\tau_i^r, rm+rt) \geq S^r(r(m-j), rm) \geq \frac{rj}{K^r} \geq \frac{j}{2K},$$

for all large r where the last inequality is due to (2.19). For those i's such that τ_i^r larger than $\bar{B}^r(rm)$, we use the trivial lower bound $S^r(\tau_i^r, rm+rt) \geq 0$. Also take the trivial lower bound that $S^r(0, rm+rt) \geq 0$. Then we have the following inequality on the $(1+p)$th moment:

$$\begin{aligned}
&\left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(rm+rt) \right\rangle \\
&\quad \leq \left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(0) \right\rangle \\
&\qquad + \sum_{j=0}^{m-1} \left\langle \left(\left(\chi - \frac{j}{2K} \right)^+ \right)^{1+p}, \frac{1}{r} \sum_{i=B^r(r(m-j-1))+1}^{B^r(r(m-j))} \delta_{v_i^r} \right\rangle \\
&\qquad + \left\langle \chi^{1+p}, \frac{1}{r} \sum_{i=B^r(rm)+1}^{B^r(rm+rt)} \delta_{v_i^r} \right\rangle.
\end{aligned} \tag{4.22}$$

Now, we consider the case where there exists an $s \in [0, rm+rt]$ such that $Z^r(s) = 0$. In this case, let $m_0 = \min\{0 \leq j < m : \text{there exists an } s \in [r(m-j-1), r(m-j)] \text{ such that } Z^r(s) = 0\}$. Pick a point $s_0 \in [r(m_0-j-1), r(m_0-j)]$ with $Z^r(s_0) = 0$. Then we can replace time 0 in (4.22) with s_0 and only consider intervals $[r(m-j-1), r(m-j)]$ with $j \leq m_0 - 1$. So we have

$$\begin{aligned}
&\left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(rm+rt) \right\rangle \\
&\quad \leq \left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(s_0) \right\rangle \\
&\qquad + \sum_{j=0}^{m_0-1} \left\langle \left(\left(\chi - \frac{j}{2K} \right)^+ \right)^{1+p}, \frac{1}{r} \sum_{i=B^r(r(m-j-1))+1}^{B^r(r(m-j))} \delta_{v_i^r} \right\rangle \\
&\qquad + \left\langle \chi^{1+p}, \frac{1}{r} \sum_{i=B^r(rm)+1}^{B^r(rm+rt)} \delta_{v_i^r} \right\rangle.
\end{aligned} \tag{4.23}$$

It is clear that the upper bound in (4.23) is less than or equal to the upper bound in (4.22). So we only need to focus on (4.22) to estimate an upper bound for the $(1+p)$th moment. By (4.20) and (4.21), for all $m \le \lfloor rT \rfloor$, $t \in [0, L]$ and all large r,

$$\begin{aligned}-rM' \le B^r(rj) \le \lambda r^2(T+1) \le r^2 M_1\\ 0 \le B^r(rj+rt) - B^r(rj)\\ \le \lambda r(L+1) + 2rM' < rL_1 \qquad \text{for all } t \in [0, L].\end{aligned}$$

It then follows from (4.22) that

$$\begin{aligned}&\left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(rm+rt)\right\rangle\\ &\quad \le \left\langle \chi^{1+p}, \frac{1}{r}\mathcal{Z}^r(0)\right\rangle\\ &\qquad + \sup_{-rM_1<n<r^2M_1} \sum_{j=0}^{m-1} \left\langle \left(\left(\chi - \frac{j}{2K}\right)^+\right)^{1+p}, \bar{\eta}^r(n, L_1)\right\rangle\\ &\qquad + \sup_{-rM_1<n<r^2M_1} \langle \chi^{1+p}, \bar{\eta}^r(n, L_1)\rangle.\end{aligned}\tag{4.24}$$

The first term on the right-hand side of the above is bounded by $1 + \langle \chi^{1+p}, \xi^* + \mu^*\rangle$ by the definition of Ω_0^r. Again, due to (4.17), condition (2.17) and Theorem 25.12 in [3], the third term on the right-hand side of the above is bounded by

$$L_1\langle \chi^{1+p}, \nu^r\rangle + 1/2 \le L_1\langle \chi^{1+p}, \nu\rangle + 1, \tag{4.25}$$

for all large r. It now only remains to deal with the second term on the right-hand side of (4.24). Let

$$\bar{F}_n^r(x) = \bar{\eta}^r(n, L_1)((x, \infty)) \qquad \text{for all } x \ge 0.$$

The summation in the second term on the right-hand side of (4.24) can be upper bounded by

$$\begin{aligned}&\frac{1}{1+p}\sum_{j=1}^{m-1}\int_{j/(2K)}^{\infty}\left(x - \frac{j}{2K}\right)^p \bar{F}_n^r(x)\,dx\\ &\quad \le \frac{2K}{1+p}\sum_{j=1}^{m-1}\int_{(j-1)/(2K)}^{j/(2K)}\int_y^{\infty}(x-y)^p \bar{F}_n^r(x)\,dx\,dy\\ &\quad \le \frac{1}{1+p}\int_0^{\infty}\int_y^{\infty}(x-y)^p \bar{F}_n^r(x)\,dx\,dy.\end{aligned}$$

Applying Fubini's theorem, the last bound in the above equals

$$\frac{1}{1+p}\int_0^\infty\int_0^x (x-y)^p \bar{F}_n^r(x)\,dy\,dx = \frac{1}{(1+p)^2}\int_0^\infty x^{1+p}\bar{F}_n^r(x)\,dx$$
$$= \frac{2+p}{(1+p)^2}\langle \chi^{2+p}, \bar{\eta}^r(n, L_1)\rangle.$$

Since the function $\chi^{2+p} \in \mathscr{V}$, it again follows from (4.17), (2.17) and Theorem 25.12 in [3] that the second term on the right-hand side of (4.24) is bounded by

$$\frac{2+p}{(1+p)^2}L_1\langle \chi^{2+p}, \nu^r\rangle + 1/2 \le \frac{2+p}{(1+p)^2}L_1\langle \chi^{2+p}, \nu\rangle + 1,$$

for all large r. The proof of this lemma is completed by summing up all these upper bounds. □

The following proposition summarizes the bound estimates in this section.

PROPOSITION 4.1. *Assume (2.14)–(2.17), (2.19)–(2.21) and (2.23). For any $\eta > 0$, there exists a constant $M > 0$ and an event $\Omega_B^r(M)$ for each index r such that*

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_B^r(M)) \ge 1 - \eta, \tag{4.26}$$

and on the event $\omega \in \Omega_B^r(M)$, we have

$$\max_{m\le \lfloor rT\rfloor}\sup_{t\in[0,L]} \bar{Q}^{r,m}(t) \le M,$$
$$\max_{m\le \lfloor rT\rfloor}\sup_{t\in[0,L]} \bar{W}^{r,m}(t) \le M,$$
$$\max_{m\le \lfloor rT\rfloor}\sup_{t\in[0,L]} \langle \chi^{1+p}, \bar{\mathcal{Q}}^{r,m}(t) + \bar{\mathcal{Z}}^{r,m}(t)\rangle \le M.$$

PROOF. The first and the third inequality follow from Lemmas 4.2 and 4.3. The second inequality follows from Proposition 2.1. □

4.3. *Compact containment.* Recall that a set $\mathbf{K} \subset \mathbf{M}$ is relatively compact if

$$\sup_{\xi\in\mathbf{K}} \xi(\mathbb{R}_+) < \infty,$$

and there exists a sequence of nested compact sets $J_n \subset \mathbb{R}_+$ such that $\bigcup J_n = \mathbb{R}_+$ 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 [17], Theorem A7.5. We establish the following relative compactness property using the bound estimates in Section 4.2.

LEMMA 4.4. *Assume (2.14)–(2.17), (2.19)–(2.21) and (2.23). Fix $T > 0$ and $L > 1$. For each $\eta > 0$ there exist a constant $M > 0$ and a relatively compact set $\mathbf{K}(M) \subset \mathbf{M}$ such that for all $\omega \in \Omega_B^r(M)$ (which is introduced in Proposition 4.1) and $r \in \mathbb{R}_+$,*

$$\bar{\mathcal{Q}}^{r,m}(\omega, t) \in \mathbf{K}(M) \quad \text{and} \quad \bar{\mathcal{Z}}^{r,m}(\omega, t) \in \mathbf{K}(M)$$

for all $t \in [0, T]$ and $m \le \lfloor rT \rfloor$.

PROOF. Let

$$\mathbf{K}(M) = \{\xi \in \mathbf{M} : \xi(\mathbb{R}_+) \le M \text{ and } \xi((n, \infty)) \le M/n^{1+p}\}.$$

Clearly, $\mathbf{K}(M)$ is a relatively compact set for any constant $M > 0$. Note that $\bar{\mathcal{Q}}^{r,m}(\omega, t)(\mathbb{R}_+)$ is bounded by M for all $m \le \lfloor rT \rfloor$, $t \in [0, T]$ and $\omega \in \Omega_B^r(M)$. By the Markov inequality, for any $t \ge 0$, $m \le \lfloor rT \rfloor$ and $\omega \in \Omega_B^r(M)$,

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

which is bounded by $\frac{M}{n^{1+p}}$ by the definition of $\Omega_B^r(M)$.

Note that $\bar{\mathcal{Z}}^{r,m}(\omega, t)(\mathbb{R}_+)$ is bounded by K^r/r by the policy constraint (2.8). By condition (2.19), $K^r/r \le K + 1$ for all large r. The same argument of $\bar{\mathcal{Q}}^{r,m}(\omega, t)$ applies for $\bar{\mathcal{Z}}^{r,m}(\omega, t)$. $\square$

4.4. *Asymptotic regularity.* A similar result as in this section was proved in [29]. However, here we consider a much longer time horizon $[0, \lfloor rT \rfloor + L]$ instead of interval $[0, T]$ in [29]. The proof of the following result use a combination of ideas in [13] and [29].

LEMMA 4.5. *Assume (2.14)–(2.23). Fix $T > 0$ and $L > 1$. For each $\varepsilon, \eta > 0$ there exists a $\kappa > 0$ (depending on ε and η) such that*

$$\liminf_{r \to \infty} \mathbb{P}^r \Big(\max_{m \le \lfloor rT \rfloor} \sup_{t \in [0, L]} \sup_{x \in \mathbb{R}_+} \bar{\mathcal{Z}}^{r,m}(t)([x, x + \kappa]) \le \varepsilon \Big) \ge 1 - \eta. \tag{4.27}$$

PROOF. To prove (4.27), it suffices to show

$$\liminf_{r \to \infty} \mathbb{P}^r \Big(\sup_{t \in [0, \lfloor rT \rfloor + L]} \sup_{x \in \mathbb{R}_+} \bar{\mathcal{Z}}^{r,0}(t)([x, x + \kappa]) \le \varepsilon \Big) \ge 1 - \eta.$$

First, we have that for any $\varepsilon, \eta > 0$, there exists a κ such that

$$\liminf_{r \to \infty} \mathbb{P}^r \Big(\sup_{x \in \mathbb{R}_+} \bar{\mathcal{Z}}^{r,0}(0)([x, x + \kappa]) \le \varepsilon/2 \Big) \ge 1 - \eta/2. \tag{4.28}$$

The proof of this inequality, which is based on (2.22), is exactly the same as the proof of (5.14) in [29], so we omit it for brevity.

Now, we need to extend this result to the interval $[0, \lfloor rT \rfloor + L]$. Denote the event in (4.28) by Ω_1^r. Let

$$\Omega_2^r(M) = \Omega_1^r \cap \Omega_E^r \cap \Omega_B^r(M) \cap \Omega_{\mathrm{GC}}^r(2\lambda T, 2M). \tag{4.29}$$

By (4.16), (4.18) and (4.26), there exists an $M > 0$ such that

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

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

For any $r > 0$, $t \in [0, \lfloor rT \rfloor + L]$, we define the time

$$t_0 = \sup\{\{s \leq t : \langle 1, \bar{\mathcal{Z}}^{r,0}(s)\rangle < \varepsilon/4\} \cup \{0\}\}$$

to be the last time before t that the fluid-scaled number of jobs in service is less than $\varepsilon/4$. Let

$$t_1 = \max\left(t_0, t - \frac{4MK}{\varepsilon}\right).$$

We have the following three cases for discussion.

If $t_1 = 0$, then by (4.28) for each $x \in \mathbb{R}_+$

$$\bar{\mathcal{Z}}^{r,0}(t_1)([x, x+\kappa] + S^r(rt_1, rt)) \leq \varepsilon/2.$$

If $t_1 = t_0 > 0$, then for each $\delta \in (0, t_1)$ there exists an $s \in (t_1 - \delta, t_1]$ such that $\bar{Z}^{r,0}(s)(\mathbb{R}_+) < \varepsilon/4$. Since we are only concerned with small ε [which should be small enough such that $\bar{Z}^{r,0}(s) < \varepsilon/4 < K^r/r$], $\bar{Q}^{r,0}(s) = 0$ by the policy constraint (2.8). Note that (2.3) implies

$$\bar{B}^{r,0}(s,t) \leq \bar{E}^{r,0}(s,t) + \bar{Q}^{r,0}(s) \qquad \text{for all } s \leq t. \tag{4.30}$$

By (4.15), we have $\bar{B}^{r,0}(s, t_1) \leq \lambda\delta + \varepsilon/4$ for all large r. For any Borel set $A \subset \mathbb{R}_+$, by the fluid scaled system dynamic equation (4.11),

$$\begin{aligned}\bar{\mathcal{Z}}^{r,0}(t_1)(A) &= \bar{\mathcal{Z}}^{r,0}(t_1)(A \cap (0,\infty)) + \bar{\mathcal{Z}}^{r,0}(t_1)(\{0\})\\ &\leq \bar{\mathcal{Z}}^{r,0}(s)(\mathbb{R}_+) + \bar{B}^{r,0}(s, t_1) \leq \varepsilon/4 + \lambda\delta + \varepsilon/4,\end{aligned}$$

which can be made smalled than $\varepsilon/2$ by choosing δ suitably small.

If $t_1 = t - \frac{2MK}{\varepsilon} > 0$, then since the sharing limit is K^r, we have $S^r(rt_1, rt) \geq \frac{4MKr}{\varepsilon K^r} \geq \frac{2M}{\varepsilon}$ for all large r. So

$$\bar{\mathcal{Z}}^{r,0}(t_1)([x, x+\kappa] + S^r(rt_1, rt)) \leq \bar{\mathcal{Z}}^{r,0}(t_1)\left(\left[\frac{2M}{\varepsilon}, \infty\right)\right) \leq \varepsilon/2,$$

where the last inequality is due to the Markov's inequality and the definition of $\Omega_B^r(M)$. To summarize, we have

$$\bar{\mathcal{Z}}^{r,0}(t_1)([x,x+\kappa]+S^r(rt_1,rt))\le\varepsilon/2. \tag{4.31}$$

By the fluid scaled stochastic dynamic equation (4.11),

$$\begin{aligned}\bar{\mathcal{Z}}^{r,0}(t)([x,x+\kappa]) &= \bar{\mathcal{Z}}^{r,0}(t_1)([x,x+\kappa]+S^r(rt_1,rt)) \\ &\quad+\frac{1}{r}\sum_{i=r\bar{B}^{r,0}(t_1)+1}^{r\bar{B}^{r,0}(t)}\delta_{v_i^r}([x,x+\kappa]+S^r(\tau_i,rt)),\end{aligned} \tag{4.32}$$

for each $x>0$. When $x=0$, we have

$$\begin{aligned}\bar{\mathcal{Z}}^{r,0}(t)([0,\kappa]) &= \bar{\mathcal{Z}}^{r,0}(t)((0,\kappa])+\bar{\mathcal{Z}}^{r,0}(t)(\{0\}) \\ &= \bar{\mathcal{Z}}^{r,0}(t_1)((0,\kappa]+S^r(rt_1,rt)) \\ &\qquad+\frac{1}{r}\sum_{i=r\bar{B}^{r,0}(t_1)+1}^{r\bar{B}^{r,0}(t)}\delta_{v_i^r}((0,\kappa]+S^r(\tau_i,rt)), \\ &\le \bar{\mathcal{Z}}^{r,0}(t_1)([0,\kappa]+S^r(rt_1,rt)) \\ &\qquad+\frac{1}{r}\sum_{i=r\bar{B}^{r,0}(t_1)+1}^{r\bar{B}^{r,0}(t)}\delta_{v_i^r}([0,\kappa]+S^r(\tau_i,rt)).\end{aligned}$$

Since all we need is an upper bound estimate, we stick with (4.32) for analysis. By the choice of t_1, the first term on the right-hand side (4.32) is always upper bounded by $\varepsilon/2$. Let I denote the second term on the right-hand side of the proceeding equation. Now it only remains to show that $I\le\varepsilon/2$.

Let $t_1<t_2<\cdots<t_N=t$ be a partition of the interval $[t_1,t]$ such that $|t_{j+1}-t_j|<\delta$ for all $j=1,\dots,N-1$, where δ is to be chosen below. Note that by the definition of t_1,

$$N\le\frac{4MK}{\delta\varepsilon}. \tag{4.33}$$

Write I as the summation

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

Recall that τ_i^r is the time that the ith job starts service, so $rt_j\le\tau_i^r\le rt_{j+1}$ for those $i\in[r\bar{B}^{r,0}(t_j)+1,r\bar{B}^{r,0}(t_{j+1})]$. This implies that

$$S^r(rt_{j+1},rt)\le S^r(\tau_i,rt)\le S^r(rt_j,rt).$$

By the definition of t_1, we have $\bar{Z}^{r,0}(s) \geq \varepsilon/4$ for all $s \in [t_1, t]$. This gives

$$S^r(rt_j, rt_{j+1}) \leq \frac{4\delta}{\varepsilon}.$$

So for any $i \in [r\bar{B}^{r,0}(t_j)+1, r\bar{B}^{r,0}(t_{j+1})]$, we have $[x, x+\kappa] + S^r(\tau_i, rt) \subseteq C_j$, where

$$C_j = \left[x + S^r(rt_{j+1}, rt), x + \kappa + S^r(rt_{j+1}, rt) + \frac{4\delta}{\varepsilon}\right].$$

Thus,

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

By (4.15), (4.30) and the definition of $\Omega_B^r(M)$, we have for all $j = 1, \ldots, N-1$

$$-rM \leq r\bar{B}^{r,0}(t_j) \leq r(\lambda r T + L + 1 + M) \leq r^2 2\lambda T,$$

$$\bar{B}^{r,0}(t_j, t_{j+1}) \leq \lambda\delta + 1 + M \leq 2M,$$

where the last inequality in each of the above bound holds because we only care about small δ and large r. Choose $\varepsilon_1 = \frac{\delta\varepsilon^2}{16MK}$. By (4.17),

$$\left|\frac{1}{r} \sum_{i=r\bar{B}^{r,0}(t_j)+1}^{r\bar{B}^{r,0}(t_{j+1})} \delta_{v_i^r}(C_j) - (\bar{B}^{r,0}(t_{j+1}) - \bar{B}^{r,0}(t_j))\nu^r(C_j)\right| \leq \varepsilon_1,$$

for all large r. This implies that

$$I \leq \sum_{j=1}^{N-1} [\bar{B}^{r,0}(t_{j+1}) - \bar{B}^{r,0}(t_j)]\nu^r(C_j) + N\varepsilon_1.$$

Let $\varepsilon_2 = \frac{\varepsilon}{8(\lambda 4MK/\varepsilon + M + 1)}$. Since $C_j^{\kappa\varepsilon_2}$ is a close interval with length $\kappa + \frac{4\delta}{\varepsilon} + \kappa\varepsilon_2$, by condition (2.18), we can choose $\kappa, \delta < 1$ small enough such that

$$\nu(C_j^{\kappa\varepsilon_2}) \leq \varepsilon_2,$$

where $C_j^{\varepsilon_2}$ is the ε_2-enlargement of C_j. By (2.16), we also have

$$\nu^r(C_j) \leq \nu(C_j^{\kappa\varepsilon_2}) + \kappa\varepsilon_2 \leq \nu(C_j^{\kappa\varepsilon_2}) + \varepsilon_2,$$

for all large enough r. Thus, we conclude that

$$\begin{aligned} I &\leq 2\varepsilon_2 \sum_{j=1}^{N-1} [\bar{B}^{r,0}(t_{j+1}) - \bar{B}^{r,0}(t_j)] + N\varepsilon_1 \\ &\leq 2\varepsilon_2[\bar{B}^{r,0}(t) - \bar{B}^{r,0}(t_1)] + N\varepsilon_1 \\ &\leq 2\varepsilon_2\left(\lambda\frac{4MK}{\varepsilon} + 1 + M\right) + N\varepsilon_1, \end{aligned}$$

where the last inequality is again due to (4.29), (4.30) and (4.33). Finally, by the choice of $\varepsilon_1, \varepsilon_2$, we have that $I \le \varepsilon/2$. □

4.5. *Oscillation bound.* Consider a càdlàg function $\zeta(\cdot)$ on a fixed interval $[0, L]$ taking values in a metric space $(\mathbf{E}, \pi)$. The *modulus of continuity* is defined to be

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

If the metric space is $\mathbb{R}$, we just use the Euclidian norm; if the space is $\mathbf{M}$ or $\mathbf{M} \times \mathbf{M}$, we use the Prohorov metric $\mathbf{d}$ defined in Section 1. We have the following bound on the oscillation of the shifted fluid scaled measure-valued processes.

LEMMA 4.6. *Assume (2.14)–(2.23). Fix $T > 0$ and $L > 1$. For each $\varepsilon, \eta > 0$ there exists a $\delta > 0$ (depending on ε and η) such that*

$$\liminf_{r\to\infty} \mathbb{P}^r\Big(\max_{m \le \lfloor rT \rfloor} \max(\mathbf{w}_L(\bar{\mathcal{Q}}^{r,m}(\cdot), \delta), \mathbf{w}_L(\bar{\mathcal{Z}}^{r,m}(\cdot), \delta)) \le \varepsilon\Big) \ge 1 - \eta. \tag{4.34}$$

The proof of this lemma, which builds on the asymptotic regularity in Lemma 4.5, uses the exactly same argument as in the proof of Lemma 5.6 based on Lemma 5.5 in [29]. We omit this proof for brevity.

Fix $T > 0$ and $L > 0$. For any sequence $\{\delta_i\}$, consider the following set:

$$\Big\{\max_{m \le \lfloor rT \rfloor} \max(\mathbf{w}_L(\bar{\mathcal{Q}}^r(\cdot), \delta_j), \mathbf{w}_L(\bar{\mathcal{Z}}^r(\cdot), \delta_j)) \le \frac{1}{j}\Big\}. \tag{4.35}$$

Denote the sequence $\{\delta_i\}$ 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 4.5 and 4.6, for any $\eta > 0$, there exists an $\mathcal{S}$ such that

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_R^r(\mathcal{S}, j)) \ge 1 - \frac{\eta/2}{2^j} \qquad \text{for } j = 1, 2, \ldots. \tag{4.36}$$

This implies that for any finite number $n \in \mathbb{N}$, 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

$$\mathbb{P}^r\left(\bigcap_{j=1}^n \Omega_R^r(\mathcal{S}, j)\right) \ge 1 - \eta \qquad \text{for all } r \ge r(n). \tag{4.37}$$

For any $n \le n'$, we have $r(n) \le r(n')$ since $\bigcap_{j=1}^{n} \Omega_R^r(\mathcal{S}, j) \supseteq \bigcap_{j=1}^{n'} \Omega_R^r(\mathcal{S}, j)$. Let

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

[By this definition, we allow $n(\cdot)$ 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 nondecreasing. Note that for any $n_0 > 0$ there exists $r_0 = r(n_0)$ such that $n(r) \ge n_0$ for all $r \ge r_0$. Thus, we have that

$$\lim_{r\to\infty} n(r) = \infty.$$

Now define

$$\Omega_R^r(\mathcal{S}) = \bigcap_{j=1}^{n(r)} \Omega_R^r(\mathcal{S}, j). \tag{4.38}$$

Note that $\Omega_R^r(\mathcal{S})$ is not empty for all large enough r [since $n(r) > 1$ for all large enough r], and in this case, $\mathbb{P}^r(\Omega_R^r(\mathcal{S})) \ge 1 - \eta$. So we conclude that

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega_R^r(\mathcal{S})) \ge 1 - \eta. \tag{4.39}$$

Denote

$$\Omega^r(M, \mathcal{S}) = \Omega_B^r(M) \cap \Omega_R^r(\mathcal{S}). \tag{4.40}$$

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 explicitly write these processes down in the form of $\mathcal{Q}^r(\omega, \cdot)$ and $\mathcal{Z}^r(\omega, \cdot)$ to indicate that they are evaluated on the sample path $\omega \in \Omega^r$. We are now ready to present the precompactness result.

THEOREM 4.1. *Assume (2.14)–(2.23). Fix $T > 0$ and $L > 1$. For each $\eta > 0$, the exists a constant $M > 0$ and an $\mathcal{S}$ such that such that*

$$\liminf_{r\to\infty} \mathbb{P}^r(\Omega^r(M, \mathcal{S})) \ge 1 - \eta. \tag{4.41}$$

Suppose $\{r_n\}_{n\in\mathbb{N}}$ is a sequence in $\mathbb{R}_+$ which goes to infinity. Any sequence of functions $\{(\bar{\mathcal{Q}}^{r_n,m_n}(\omega_n, \cdot), \bar{\mathcal{Z}}^{r_n,m_n}(\omega_n, \cdot))\}_{n\in\mathbb{N}}$ with $\omega_n \in \Omega^{r_n}(M, \mathcal{S})$ and $m_n \le \lfloor r_n T\rfloor$ for each $n \in \mathbb{N}$ has a subsequence $\{(\bar{\mathcal{Q}}^{r_{n_i},m_{n_i}}(\omega_{n_i}, \cdot), \bar{\mathcal{Z}}^{r_{n_i},m_{n_i}}(\omega_{n_i}, \cdot))\}_{i\in\mathbb{N}}$ such that

$$\upsilon_L[(\bar{\mathcal{Q}}^{r_{n_i},m_{n_i}}(\omega_{n_i}, \cdot), \bar{\mathcal{Z}}^{r_{n_i},m_{n_i}}(\omega_{n_i}, \cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] \to 0 \qquad \text{as } i \to \infty,$$

for some process $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ which is continuous, where υ_L is the uniform metric defined in (1.1).

PROOF. For a fixed $\eta > 0$, pick an $M > 0$ that satisfies (4.26) and construct an $\mathcal{S}$ so that it satisfies (4.36). Define $\Omega^r(M, \mathcal{S})$ via (4.40). The probability inequality (4.41) follows immediately from (4.26) and (4.39). The space $\mathbf{M} \times \mathbf{M}$ endowed with the metric $\mathbf{d}$ (defined in Section 1.1) is complete. Lemma 4.4 verifies condition (a) in Theorem 3.6.3 of [11]. For any $\varepsilon > 0$ there exists a j_0 such that $1/j < \varepsilon$ for all $j > j_0$. By (4.35) and (4.38), we have that when $\delta \le \delta_{j_0}$ and $r > r(j_0)$, where δ_{j_0} is specified in $\mathcal{S}$ and $r(n)$ is defined in (4.37),

$$\max(\mathbf{w}_T(\bar{\mathcal{Q}}^{r,m}(\omega^r, \cdot), \delta), \mathbf{w}_T(\bar{\mathcal{Z}}^{r,m}(\omega^r, \cdot), \delta)) \le \varepsilon, \tag{4.42}$$

for any $\omega^r \in \Omega^r(M, \mathcal{S})$ and $m \le \lfloor rT \rfloor$. This verifies condition (b) in Theorem 3.6.3 of [11]. So the sequence $\{(\bar{\mathcal{Q}}^{r_n,m_n}(\omega^{r_n}, \cdot), \bar{\mathcal{Z}}^{r_n,m_n}(\omega^{r_n}, \cdot))\}_{n\in\mathbb{N}}$ is precompact in the space $\mathbf{D}([0,T], \mathbf{M} \times \mathbf{M})$ 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 (4.42). So convergence in the Skorohod J_1-topology is the same as convergence in the uniform metric defined in Section 1.1. $\square$

5. State-space collapse. In this section, we establish the state-space collapse (Theorem 2.2). The task is divided into the following steps: we first show that the limits in Theorem 4.1, which called fluid limits, are fluid model solutions; the set of fluid limits is "rich" in the sense that itself and the set of shifted fluid scaled process mutually approximates each other (Lemmas 5.1 and 5.3); the proof of the state-space collapse result is finally presented based on the richness of fluid limits and the properties of fluid model solution (Theorems 3.1 and 3.2).

5.1. *Fluid limits.* Let $\mathscr{D}_L(M, \mathcal{S})$ denote the set of fluid limits of all convergent subsequences of sequences in Theorem 4.1. It is then quite clear that we have the following property.

LEMMA 5.1. *Assume (2.14)–(2.23). The set of fluid limits $\mathscr{D}_L(M, \mathcal{S})$ is nonempty. Pick an element $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$, for any $\varepsilon > 0$ and $r_0 \in \mathbb{R}_+$, there exists an $r \ge r_0$, $m \le \lfloor rT \rfloor$ and $\omega \in \Omega^r(M, \mathcal{S})$ such that*

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,m}(\omega, \cdot), \bar{\mathcal{Z}}^{r,m}(\omega, \cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] \le \varepsilon.$$

Roughly speaking, this lemma says that any element in $\mathscr{D}_L(M, \mathcal{S})$ can be approximated by a shifted fluid scaled process of the rth system evaluated at some sample path in $\Omega^r(M, \mathcal{S})$ with arbitrarily large index r. This helps prove the following property of the fluid limits.

Fix a constant $0 < q < p$, where p is the same one as in (2.17) and (2.21). Recall the subset $\mathscr{I}^q_{3M}$ of all valid initial conditions defined in (3.21).

LEMMA 5.2. *Assume (2.14)–(2.23). Fix $L > 0$ and $0 < q < p$. Any element $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ is a critically loaded fluid model solution with initial condition belongs to $\mathscr{I}_{3M}^q$.*

PROOF. We first show that the initial condition $(\tilde{\mathcal{Q}}(0), \tilde{\mathcal{Z}}(0)) \in \mathscr{I}_{3M}^q$. By the definition of the fluid limit, there exists a subsequence

$$(\bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0), \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)) \to (\tilde{\mathcal{Q}}(0), \tilde{\mathcal{Z}}(0)) \qquad \text{as } i \to \infty,$$

where the above convergence is in the Prohorov metric. By Proposition 4.1 and the LPS policy, we have

$$\langle 1, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle < M + K + 1,$$
$$\langle \chi^{1+p}, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0), \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle < M,$$

for all large r_i. This implies that for any $0 \le q < p$,

$$\begin{aligned}
&\langle \chi^{1+q}, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle \\
&\quad \le \langle 1, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle + \langle \chi^{1+p}, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle \\
&\quad < 2M + K + 1.
\end{aligned}$$

By the corollary of Theorem 25.12 in [3], we have that for any $0 \le q < p$,

$$\langle \chi^{1+q}, \bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, 0) + \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, 0)\rangle \to \langle \chi^{1+q}, \tilde{\mathcal{Q}}(0) + \tilde{\mathcal{Z}}(0)\rangle \qquad \text{as } i \to \infty.$$

Since we can take M big enough such that $M > K + 1$, this implies that $\langle \chi^{1+q}, \tilde{\mathcal{Q}}(0) + \tilde{\mathcal{Z}}(0)\rangle < 3M$ and $\langle \chi, \tilde{\mathcal{Q}}(0) + \tilde{\mathcal{Z}}(0)\rangle < 3M$, which yields the result.

By Lemma 5.1, any fluid limit $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ can be approximated by a shifted fluid scaled process of the rth system evaluated at some sample path in $\Omega^r(M, \mathcal{S})$ with arbitrarily large index $r \in \mathbb{R}_+$; the state descriptor of the rth system satisfies the stochastic dynamic equations (2.5) and (2.6). It then follows from the same argument as in Lemmas 6.1 and 6.2 in [29] that each fluid limit satisfies the fluid model equations (3.12) and (3.13) and constraints (3.14)–(3.16). In fact, [29] is more general in the sense that the traffic intensity is allowed to be any positive number instead of being 1 as required in this paper. □

5.2. *Uniform approximation.* Lemma 5.3 in the following is analogous to Lemma 4.1 in [6]. In contrast to Lemma 5.1 above, this lemma says that any shifted fluid scaled process of the rth system evaluated at some sample path in $\Omega^r(M, \mathcal{S})$ with index r large enough can be approximated by some element in $\mathscr{D}_L(M, \mathcal{S})$, which has been proved to be a fluid model solution in Lemma 5.2. This result will help prove the state-space collapse result for diffusion scaled processes.

LEMMA 5.3. *Assume (2.14)–(2.23). For each $\varepsilon > 0$, there exists an $r_0 \in \mathbb{R}_+$ such that for any $r \geq r_0$, $m \leq \lfloor rT \rfloor$ and $\omega \in \Omega^r(M, \mathcal{S})$, we can find a $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ satisfying*

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,m}(\omega, \cdot), \bar{\mathcal{Z}}^{r,m}(\omega, \cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] < \varepsilon.$$

PROOF. Assume it is not true. Then there exists an $\varepsilon > 0$ such that for any natural number i there exist an $r_i > i$, $m_i \in \lfloor rT \rfloor$ and $\omega_i \in \Omega^r(M, \mathcal{S})$ such that

$$\upsilon_L[(\bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, \cdot), \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, \cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] \geq \varepsilon,$$

for all $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$. However, by Theorem 4.1, the sequence

$$\{(\bar{\mathcal{Q}}^{r_i,m_i}(\omega_i, \cdot), \bar{\mathcal{Z}}^{r_i,m_i}(\omega_i, \cdot))\}_{i=0}^{\infty}$$

contains a convergent subsequence, the limit of which must be in $\mathscr{D}_L(M, \mathcal{S})$. This is a contradiction. □

5.3. *Proof of state-space collapse.* With all the preparation, we finally present the proof of state-space collapse.

PROOF OF THEOREM 2.2. By (4.41), it suffices to show that for each $\varepsilon > 0$, there exists an r_0 such that when $r > r_0$,

$$\sup_{\omega \in \Omega^r(M,\mathcal{S})} \sup_{t \in [0,T]} \mathbf{d}[(\hat{\mathcal{Q}}^r(\omega, t), \hat{\mathcal{Z}}^r(\omega, t)), \Delta_{K,\lambda} \hat{W}^r(\omega, t)] < \varepsilon. \tag{5.1}$$

In the following, we fix $r > r_0$ and $\omega \in \Omega^r(M, \mathcal{S})$. By Lemma 5.2, any $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ is a critically loaded fluid model solution with initial condition $(\xi, \mu) \in \mathscr{I}_{3M}^q$. Denote

$$\tilde{W}(\cdot) = \langle \chi, \tilde{\mathcal{Q}}(\cdot) + \tilde{\mathcal{Z}}(\cdot) \rangle.$$

It follows from the workload conservation property (3.20) that $\tilde{W}(\cdot) \equiv \langle \chi, \xi + \mu \rangle$. By Theorem 3.1, there exists an $L^* > 0$ such that when $s > L^*$,

$$\mathbf{d}[(\tilde{\mathcal{Q}}(s), \tilde{\mathcal{Z}}(s)), \Delta_{K,\nu} \tilde{W}(s)] < \varepsilon/3, \tag{5.2}$$

for all $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$. Now, fix a constant $L > L^* + 1$. Note that

$$[0, r^2 T] \subset [0, rL^*] \cup \bigcup_{m=0}^{\lfloor rT \rfloor} [r(m + L^*), r(m + L)].$$

By the definition of diffusion and shifted fluid scaling, to show (5.1) it suffices to show

$$\max_{m \leq \lfloor rT \rfloor} \sup_{s \in [L^*, L]} \mathbf{d}[(\bar{\mathcal{Q}}^{r,m}(\omega, s), \bar{\mathcal{Z}}^{r,m}(\omega, s)), \Delta_{K,\lambda} \bar{W}^{r,m}(\omega, s)] < \varepsilon, \tag{5.3}$$

$$\sup_{s \in [0, L^*]} \mathbf{d}[(\bar{\mathcal{Q}}^{r,0}(\omega, s), \bar{\mathcal{Z}}^{r,0}(\omega, s)), \Delta_{K,\lambda} \bar{W}^{r,0}(\omega, s)] < \varepsilon. \tag{5.4}$$

We first prove (5.3). Fix an $m \le \lfloor rT \rfloor$. By Lemma 5.3, for any $\varepsilon' > 0$, there exists a $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ (depending on r, m and ω) such that

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,m}(\omega, \cdot), \bar{\mathcal{Z}}^{r,m}(\omega, \cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] < \varepsilon'. \tag{5.5}$$

By the definition of $\Omega^r(M, \mathcal{S})$ and Proposition 4.1, following the same proof as in Lemma 5.2, we have that for each fixed $0 < q < p$,

$$\langle \chi^{1+q}, \tilde{\mathcal{Q}}(t) + \tilde{\mathcal{Z}}(t) \rangle < 3M,$$
$$\langle \chi^{1+q}, \bar{\mathcal{Q}}^{r,m}(\omega, t) + \bar{\mathcal{Z}}^{r,m}(\omega, t) \rangle < 3M,$$

for all $t \in [0, L]$. It then follows from Lemma C.2 and by taking ε' small enough that

$$\sup_{t \in [0,L]} |\tilde{W}(t) - \bar{W}^{r,m}(\omega, t)| < \frac{\varepsilon}{3 \max(1/\beta, 1/\beta_e)}. \tag{5.6}$$

Note that for any real numbers w_1, w_2, by the definition of the lifting map $\Delta_{K,\nu}$ and the metric $\mathbf{d}$, we have

$$\begin{aligned} &\mathbf{d}[\Delta_{K,\nu} w_1, \Delta_{K,\nu} w_2] \\ &\quad < \max\biggl(\mathbf{d}\biggl[\frac{(w_1 - K\beta_e)^+}{\beta}\nu, \frac{(w_2 - K\beta_e)^+}{\beta}\nu\biggr], \\ &\qquad\qquad \mathbf{d}\biggl[\frac{w_1 \wedge K\beta_e}{\beta_e}\nu_e, \frac{w_2 \wedge K\beta_e}{\beta_e}\nu_e\biggr]\biggr). \end{aligned}$$

It is clear that for any $a, b \ge 0$ and Borel set $A \subseteq \mathbb{R}$, we have that $a\nu(A) \le b\nu(A) + |b - a| \le b\nu(A^{|b-a|}) + |b - a|$, where $A^{|b-a|}$ is the $|b-a|$-enlargement of A. Similarly, we have $b\nu(A) \le b\nu(A^{|b-a|}) + |b - a|$. So $\mathbf{d}[a\nu, b\nu] \le |b - a|$. This implies that

$$\begin{aligned} \mathbf{d}\biggl[\frac{(w_1 - K\beta_e)^+}{\beta}\nu, \frac{(w_2 - K\beta_e)^+}{\beta}\nu\biggr] &\le \biggl|\frac{(w_1 - K\beta_e)^+}{\beta} - \frac{(w_2 - K\beta_e)^+}{\beta}\biggr| \\ &\le \frac{1}{\beta}|w_1 - w_2|. \end{aligned}$$

Following the same argument, we have

$$\mathbf{d}\biggl[\frac{w_1 \wedge K\beta_e}{\beta_e}\nu_e, \frac{w_2 \wedge K\beta_e}{\beta_e}\nu_e\biggr] \le \biggl|\frac{w_1 \wedge K\beta_e}{\beta_e} - \frac{w_2 \wedge K\beta_e}{\beta_e}\biggr| \le \frac{1}{\beta_e}|w_1 - w_2|.$$

Thus, we conclude that

$$\mathbf{d}[\Delta_{K,\nu} w_1, \Delta_{K,\nu} w_2] < \max\biggl(\frac{1}{\beta}, \frac{1}{\beta_e}\biggr)|w_1 - w_2|. \tag{5.7}$$

So (5.3) follows from (5.2) and (5.5)–(5.7).

It now remains to show (5.4). By Lemma 5.3, for any $\varepsilon' > 0$, there exists a $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot)) \in \mathscr{D}_L(M, \mathcal{S})$ (depending on r and ω) such that

$$\upsilon_L[(\bar{\mathcal{Q}}^{r,0}(\omega,\cdot), \bar{\mathcal{Z}}^{r,0}(\omega,\cdot)), (\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))] < \varepsilon'. \tag{5.8}$$

By conditions (2.20) and (2.24), we have that

$$(\tilde{\mathcal{Q}}(0), \tilde{\mathcal{Z}}(0)) = \Delta_{K,\nu}\tilde{W}(0).$$

In other words, the initial condition $(\tilde{\mathcal{Q}}(0), \tilde{\mathcal{Z}}(0))$ is an equilibrium state. Since $(\tilde{\mathcal{Q}}(\cdot), \tilde{\mathcal{Z}}(\cdot))$ is a fluid model solution, by Theorem 3.2,

$$(\tilde{\mathcal{Q}}(t), \tilde{\mathcal{Z}}(t)) = \Delta_{K,\nu}\tilde{W}(t) \qquad \text{for all } t \ge 0.$$

So (5.4) follows immediately from (5.6)–(5.8) and the above equation. $\square$

APPENDIX A: AN INTEGRATION BY PARTS FORMULA FOR LEBESGUE–STIELTJES INTEGRAL

The following lemma is used in the derivation of (3.25) and in the proof of Lemma 3.3. We do not require the continuity of distribution function F.

LEMMA A.1. *Suppose that F is a probability distribution function with $F(0) = 0$, and $q \in \mathbf{D}([0,\infty), \mathbb{R})$ has bounded variation on $[0,b]$ for each $b > 0$. For any $u > 0$,*

$$\int_{[0,u]} [1 - F(u-v)]\, dq(v) = q(u) - [1 - F(u)]q(0) - \int_{[0,u]} q(u-v)\, dF(v).$$

PROOF. Let $u > 0$ be fixed and let $I = [0,u]$. Define $f(v) = 1 - F(u-v)$. Then f is a left continuous function in on $(-\infty, u]$. Clearly, both f and q are functions with bounded variation on the interval I. Let S denote the set of points in I where both f and q are discontinuous. According to Theorem 6.2.2 in [7],

$$\begin{aligned} \int_I f\, dq + \int_I q\, df &= f(u^+)q(u^+) - f(0^-)q(0^-) + \sum_{a\in S} A(a) \\ &= q(u) - [1 - F(u)]q(0) + \sum_{a \in S} A(a), \end{aligned} \tag{A.1}$$

where

$$\begin{aligned} A(a) = {} & [f(a) - \tfrac{1}{2}(f(a^+) + f(a^-))](q(a^+) - q(a^-)) \\ & + [q(a) - \tfrac{1}{2}(q(a^+) + q(a^-))](f(a^+) - f(a^-)). \end{aligned}$$

Since f is continuous on the left and q is continuous on the right at all $a \in S$, then

$$A(a) = [-\tfrac{1}{2}(f(a^+) - f(a))][q(a) - q(a^-)] + [\tfrac{1}{2}(q(a) - q(a^-))][f(a^+) - f(a)] = 0. \tag{A.2}$$

Now the lemma follows from (A.1), (A.2) and

$$\int_I q\,df = \int_{[0,u]} q(v)\,dF(u-v) = \int_{[0,u]} q(u-v)\,dF(v).$$

□

APPENDIX B: A KEY RENEWAL THEOREM WITH UNIFORM CONVERGENCE

The following result is similar as the key renewal theorem. But the convergence is shown to be uniform on a set of functions $\mathscr{H}$ as specified below.

Lemma B.1. *Assume that each $h \in \mathscr{H}$ is nonnegative and nonincreasing. Assume that*

$$M_1 = \sup_{h\in\mathscr{H}} h(0) < \infty, \tag{B.1}$$

$$\lim_{x\to\infty} \sup_{h\in\mathscr{H}} \int_x^\infty h(y)\,dy = 0. \tag{B.2}$$

Let U be the renewal function associated with a nonlattice inter-renewal distribution with finite mean β. Then

$$\lim_{x\to\infty} \sup_{h\in\mathscr{H}} \left| U * h(x) - \frac{1}{\beta}\int_0^\infty h(y)\,dy \right| = 0. \tag{B.3}$$

Proof. Conditions (B.1) and (B.2) imply that

$$M_2 \equiv \sup_{h\in\mathscr{H}} \int_0^\infty h(y)\,dy < \infty.$$

Let $\delta > 0$ and $\varepsilon > 0$ be arbitrary positive numbers in $(0,1)$. By (B.2), there exists $N = N(\delta,\varepsilon)$ such that for each $h \in \mathscr{H}$

$$\int_{N\delta}^\infty h(y)\,dy < \delta\varepsilon.$$

Furthermore, by the Blackwell theorem, there exists x^* such that for each $x \geq x^*$ and each $k = 0, \ldots, N$,

$$\frac{\delta}{\beta} - \delta\varepsilon < U(x - (k+1)\delta) - U(x - k\delta) < \frac{\delta}{\beta} + \delta\varepsilon.$$

Let $d_k(x) = U(x - k\delta) - U(x - (k+1)\delta)$ for all $k \geq 0$. Here, we take the convention that $U(x) = 0$ for all $x < 0$. Define

$$\bar{h}^\delta(x) = \sum_{k=0}^{\infty} h(k\delta) 1_{\{k\delta \leq x < (k+1)\delta\}}.$$

Clearly, $h(x) \leq \bar{h}^\delta(x)$ for $x \geq 0$. So for $x \geq x^*$ and $h \in \mathscr{H}$, we have

$$\begin{aligned}
U * h(x) \leq U * \bar{h}^\delta(x) &= \sum_{k=0}^{\infty} h(k\delta) d_k(t) \\
&= \sum_{k=0}^{N} h(k\delta) d_k(t) + \sum_{k=N+1}^{\infty} h(k\delta) d_k(t) \\
&\leq \sum_{k=0}^{N} h(k\delta) \left(\frac{\delta}{\beta} + \varepsilon\delta\right) + U(\delta) \sum_{k=N+1}^{\infty} h(k\delta) \\
&\leq \left(\delta h(0) + \int_0^\infty h(y)\,dy\right)\left(\frac{1}{\beta} + \varepsilon\right) + U(\delta)\frac{1}{\delta}\int_{N\delta}^\infty h(y)\,dy \\
&\leq \left(\delta h(0) + \int_0^\infty h(y)\,dy\right)\left(\frac{1}{\beta} + \varepsilon\right) + U(1)\varepsilon \\
&\leq \frac{1}{\beta}\int_0^\infty h(y)\,dy + \delta M_1\left(\frac{1}{\beta} + \varepsilon\right) + M_2\varepsilon + U(1)\varepsilon.
\end{aligned}$$

Define

$$\underline{h}^\delta(t) = \sum_{k=0}^{\infty} h((k+1)\delta) 1_{\{k\delta \leq t < (k+1)\delta\}}.$$

Clearly, $h(x) \geq \underline{h}^\delta(x)$ for $x \geq 0$. So for $x \geq x^*$ and $h \in \mathscr{H}$, we have

$$\begin{aligned}
U * h(x) \geq U * \underline{h}^\delta(x) &= \sum_{k=0}^{\infty} h((k+1)\delta) d_k(t) \\
&= \sum_{k=0}^{N} h((k+1)\delta) d_k(t) + \sum_{k=N+1}^{\infty} h((k+1)\delta) d_k(t)
\end{aligned}$$

$$
\begin{aligned}
&\geq \sum_{k=0}^{N} h((k+1)\delta)\left(\frac{\delta}{\beta} - \varepsilon\delta\right) \\
&= \left(\sum_{k=0}^{\infty} h(k\delta) - \delta h(0) - \sum_{k=N+2}^{\infty} h(k\delta)\right)\left(\frac{\delta}{\beta} - \varepsilon\delta\right) \\
&\geq \left(\int_0^\infty h(y)\,dy - \int_{(N+1)\delta}^\infty h(y)\,dy\right)\left(\frac{1}{\beta} - \varepsilon\right) - \delta h(0)\left(\frac{\delta}{\beta} - \varepsilon\delta\right) \\
&\geq \frac{1}{\beta}\int_0^\infty h(y)\,dy - M_2\varepsilon - \delta\varepsilon\left(\frac{1}{\beta} - \varepsilon\right) - \delta^2 M_1\left(\frac{1}{\beta} - \varepsilon\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\limsup_{x\to\infty} \sup_{h\in\mathscr{H}} \left| U * h(x) - \frac{1}{\beta}\int_0^\infty h(y)\,dy \right| \\
&\qquad \leq \delta M_1\left(\frac{1}{\beta} + \varepsilon\right) + 2M_2\varepsilon + U(1)\varepsilon + \delta\varepsilon\left(\frac{1}{\beta} - \varepsilon\right) \\
&\qquad\quad + \delta^2 M_1\left(\frac{1}{\beta} - \varepsilon\right).
\end{aligned}
$$

Because $\delta > 0$ and $\varepsilon > 0$ can be arbitrarily small, we have

$$
\lim_{x\to\infty} \sup_{h\in\mathscr{H}} \left| U * h(x) - \frac{1}{\beta}\int_0^\infty h(y)\,dy \right| = 0. \qquad \square
$$

APPENDIX C: SOME RESULTS ON THE PROHOROV METRIC

Lemma C.1 is applied in Section 3.3, and Lemma C.2 is applied in Section 5.3. Since we could not find these results in the literature, we include them here for completeness.

LEMMA C.1. *Let μ and μ_1 be finite Borel measures on $[0,\infty)$. Denote $A_y = (y,\infty)$ for all $y \geq 0$. Let $M = \max(\langle \chi, \mu\rangle, \langle \chi, \mu_1\rangle)$. For all $0 < \varepsilon < 1$ if*

$$
\sup_{y\geq 0} |\mu(A_y) - \mu_1(A_y)| < \varepsilon, \tag{C.1}
$$

then

$$
\mathbf{d}[\mu, \mu_1] < (M+2)\varepsilon^{1/3}.
$$

PROOF. Let α, β be positive constants to be determined later. Note that

$$
\mu((\varepsilon^{-\alpha}, \infty)) \leq M\varepsilon^{\alpha}.
$$

For any real number a, denote $I_a = (a, a+\varepsilon^\beta]$. Condition (C.1) implies that

$$\sup_{a\in\mathbb{R}} |\mu(I_a) - \mu_1(I_a)| < 2\varepsilon.$$

For any Borel set $A \subset [0,\infty)$, there exist $a_1, \ldots, a_N$ such that

$$A \cap [0, \varepsilon^{-\alpha}] \subset \bigcup_{i=1}^N I_{a_i},$$

and $I_{a_i} \cap I_{a_j} = \varnothing$ for all $i \neq j$, and $I_{a_i} \cap A \neq \varnothing$ for all i. These conditions imply that

$$N \leq \varepsilon^{-\alpha-\beta}$$

and

$$\bigcup_{i=1}^N I_{a_i} \subset A^{\varepsilon^\beta},$$

where A^{ε^β} is the ε^β-enlargement of the set defined in Section 1.1. So we have

$$\begin{aligned}
\mu(A) &\leq \mu(A \cap [0, \varepsilon^{-\alpha}]) + \mu(A \cap (\varepsilon^{-\alpha}, \infty)) \\
&\leq \mu\left(\bigcup_{i=1}^N I_{a_i}\right) + M\varepsilon^\alpha \\
&< \mu_1\left(\bigcup_{i=1}^N I_{a_i}\right) + N2\varepsilon + M\varepsilon^\alpha \\
&\leq \mu_1(A^{\varepsilon^\beta}) + 2\varepsilon^{1-\alpha-\beta} + M\varepsilon^\alpha.
\end{aligned}$$

Now choose $\alpha = \beta = 1/3$ to obtain

$$\mu(A) < \mu_1(A^{(M+2)\varepsilon^{1/3}}) + (M+2)\varepsilon^{1/3}.$$

Exchanging the position of μ and μ_1 in the above argument, we have

$$\mu_1(A) < \mu(A^{(M+2)\varepsilon^{1/3}}) + (M+2)\varepsilon^{1/3}.$$

This completes the proof. □

LEMMA C.2. *Suppose μ_1 and μ are finite Borel measures on $\mathbb{R}_+$ satisfying*

$$\mathbf{d}[\mu_1, \mu] < \varepsilon < 1, \tag{C.2}$$

and $\langle \chi^{1+q}, \mu_1\rangle < M$, $\langle \chi^{1+q}, \mu\rangle < M$ for some positive constants q and M, then

$$|\langle \chi, \mu_1\rangle - \langle \chi, \mu\rangle| \leq \varepsilon^{1/2} + \frac{2M}{q}\varepsilon^{q/2}.$$

PROOF. By Markov inequality, $\mu_1(A_x) \le \frac{M}{x^{1+q}}$ and $\mu(A_x) \le \frac{M}{x^{1+q}}$ for all $x \ge 0$. For any $C > 0$, we have the following inequality:

$$\begin{aligned}|\langle \chi, \mu_1 \rangle - \langle \chi, \mu \rangle| &\le \int_0^C |\mu_1(A_x) - \mu(A_x)|\, dx \\ &\quad + \int_C^\infty \mu_1(A_x)\, dx + \int_C^\infty \mu(A_x)\, dx \\ &\le C\varepsilon + 2\int_C^\infty \frac{M}{x^{1+q}}\, dx \\ &= C\varepsilon + 2M\frac{1}{qC^q}.\end{aligned}$$

The result follows by letting $C = \varepsilon^{-1/2}$. □

APPENDIX D: GLIVENKO–CANTELLI ESTIMATE

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; those with $i \le 0$ correspond to the service requirement of initial jobs waiting in the buffer. For any $n \in \mathbb{Z}$ and $l \in \mathbb{R}_+$, define

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

The objective of this section is to obtain the *Glivenko–Cantelli estimate*, Lemma D.1 below, for $\bar{\eta}^r(n, l)$. Very similar result was shown in Lemma 4.7 [13]. For completeness, the proof which follows the one in [13] is provided here.

To present the result, we introduce some notions from empirical process theory. Our primary references are [13] and [25].

A collection $\mathcal{C}$ of subsets of $\mathbb{R}^2$ shatters an n-point subset $\{x_1, \dots, x_n\} \subset \mathbb{R}_+$ if the collection $\{\mathcal{C} \cap \{x_1, \dots, x_n\} : C \in \mathcal{C}\}$ has cardinality 2^n. In this case, we say that $\mathcal{C}$ picks out all subsets of $\{x_1, \dots, x_n\}$. The *Vapnik–Červonenkis index* (*VC-index*) of $\mathcal{C}$ is

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

where the minimum of the empty set equals infinity. The collection $\mathcal{C}$ is a *Vapnik–Červonenkis class* (*VC-class*) if it has finite VC-index. Let $\mathscr{V}$ be a family of Borel measurable functions $f : R_+ \to \mathbb{R}$. We call $\mathscr{V}$ a VC-class if the collection of subgraphs $\{\{(x, y) : y < f(x)\} : f \in \mathscr{V}\}$ is a VC-class of sets in $\mathbb{R}^2$.

We call a family of functions $\mathscr{V}$ a *Borel measurable class* if, for each $n \in \mathbb{N}$ and $(e_1, \dots, e_n) \in \{-1, 1\}^n$, the map

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

is Borel measurable on $\mathbb{R}_+^n$. The condition requires that, for all $\delta > 0$ and $r \in \mathbb{R}_+$, the families $\mathscr{V}_\delta^r = \{f - g : f, g \in \mathscr{V}, \|f - g\|_{\nu^r,2} < \delta\}$ and $\mathscr{V}_\infty^2 = \{(f - g)^2 : f, g \in \mathscr{V}\}$ are Borel measurable, where

$$\|f\|_{\nu^r,2} = \langle |f|^2, \nu \rangle^{1/2}$$

denotes the $L_2(\nu^r)$-norm.

We call a Borel measurable function $\bar{f} : R_+ \to \mathbb{R}$ an envelope function for $\mathscr{V}$ if any element in $\mathscr{V}$ is bounded by $\bar{f}$. A VC-class with an envelop function satisfies a very useful entropy bound. Let Γ be the set of finitely discrete probability measures γ on $\mathbb{R}_+$ such that $\|\bar{f}\|_{\gamma,2} > 0$. For any Borel measurable function $f : \mathbb{R}_+ \to \mathbb{R}$ satisfying $\|f\|_{\gamma,2} < \infty$, let $B_f(\varepsilon) = \{g \in \mathscr{V} : \|g - f\|_{\nu^r,2} < \varepsilon\}$ denote the $L_2(\nu^r)$-ball in $\mathscr{V}$, centered at f with radius ε. For a family of functions $\mathscr{V}$, $N(\varepsilon, \mathscr{V}, L_2(\gamma))$ is the smallest number of balls $B_f(\varepsilon)$ needed to cover $\mathscr{V}$. Then $\mathscr{V}$ satisfies

$$\text{(D.2)} \qquad \int_0^\infty \sup_{\gamma \in \Gamma} \sqrt{\log N(\varepsilon \|\bar{f}\|_{\gamma,2}, \mathscr{V}, L_2(\gamma))}\, d\varepsilon < \infty;$$

see Definition 2.1.5, (2.5.1) and Theorem 2.6.7 in [25].

LEMMA D.1. *Let $\mathscr{V}$ be a VC-class of Borel measurable functions such that $\mathscr{V}_\infty^2$ and $\mathscr{V}_\delta^r$ are Borel measurable classes for all $r \in \mathbb{R}_+$ and $\delta > 0$. Assume there exists an envelop function $\bar{f}$ of $\mathscr{V}$ such that*

$$\text{(D.3)} \qquad \lim_{N \to \infty} \sup_{r \in \mathbb{R}_+} \langle \bar{f}^2 1_{\bar{f} > N}, \nu^r \rangle = 0.$$

Fix constants $M_1, L_1 > 0$. For all $\varepsilon, \varepsilon' > 0$,

$$\text{(D.4)} \qquad \limsup_{r \to \infty} \mathbb{P}^r \Big(\max_{-rM_1 < n < r^2 M_1} \sup_{l \in [0, L_1]} \sup_{f \in \mathscr{V}} |\langle f, \bar{\eta}^r(n, l) \rangle - l \langle f, \nu^r(A_x) \rangle| > \varepsilon' \Big) < \varepsilon.$$

REMARK D.1. To apply the lemma in this paper, we take

$$\text{(D.5)} \qquad \mathscr{V} = \{1_C : C \in \mathcal{C}\} \cup \{\chi^{1+p}, \chi^{2+p}\},$$

where $\mathcal{C} = \{[y, \infty) : y \in \mathbb{R}_+\} \cup \{(y, \infty) : y \in \mathbb{R}_+\}$ and p is the same one as in condition (2.17). It is very easy to see that $\mathscr{V}$ is a VC-class. Note that both $\mathscr{V}_\delta^r$ and $\mathscr{V}_\infty^r$ are subsets of functions of the form $1_{(a,b)}$ (or $1_{(a,b]}$, $1_{[a,b)}$, $1_{[a,b]}$).

So the supreme over these two families will be the same as supreme over all a, b in subsets of $\mathbb{R}_+$, which will be the same as over all a, b in subsets of Q. Borel measurability is preserved when take supreme over a countable set. It is also clear that

$$\bar{f}(x) = \begin{cases} 1, & x < 1, \\ x^{2+p}, & x \geq 1, \end{cases}$$

is an envelop function. Condition (2.17) implies (D.3).

To better structure the proof, we present the following auxiliary lemma.

LEMMA D.2. *For $n \in \mathbb{Z}$ and $k \in \mathbb{N}$, define*

$$\xi^r_{n,k} = \frac{1}{\sqrt{k}} \sum_{i=n+1}^{n+k} (\delta_{v^r_i} - \nu^r). \tag{D.6}$$

Then for any $q > 1$, $y > 2$ and $n \in \mathbb{Z}$ there exists $M_q < \infty$ and k_0 such that $k \geq k_0$ implies

$$\sup_r \mathbb{P}^r \Big(\sup_{f \in \mathscr{V}} \langle f, \xi^r_{n,k} \rangle > y \Big) < \frac{M_q}{y^q}. \tag{D.7}$$

The constant M_q does not depend on y.

PROOF. Let us first fix $n = 0$ and look at $\xi^r_{0,k}$ which will be denoted by ξ^r_k for simplicity. The property (D.3) of the envelop function $\bar{f}$ and the uniform entropy bound (D.2), together with the sets $\mathscr{V}^r_\delta$ and $\mathscr{V}^r_\infty$ being Borel measurable, imply that $\mathscr{V}$ is Donsker and pre-Gaussian uniformly in ν^r, $r \in \mathbb{R}_+$. (See Theorem 2.8.3 in [25].)

Let $l^\infty(\mathscr{V})$ be the space of all probability measures on $\mathbb{R}_+$ equipped with norm $\|\cdot\|_{\mathscr{V}} = \sup_{f \in \mathscr{V}} \langle f, \cdot \rangle$. $\mathscr{V}$ being Donsker uniformly in ν^r means that ξ^r_k converges weakly as $n \to \infty$ in $l^\infty(\mathscr{V})$ to a tight, Borel measurable version of the Brownian bridge ξ^r uniformly for all ν^r. According to Chapter 1.12 in [25], this is equivalent to

$$\sup_{h \in \mathrm{BL}_1} |\mathbb{E}^r h(\xi^r_k) - \mathbb{E} h(\xi^r)| \to 0. \tag{D.8}$$

uniformly for all ν^r, where BL_1 is the set of functions $h : l^\infty(\mathscr{V}) \to \mathbb{R}$ which are uniformly bounded by 1 and satisfy $|h(z_1) - h(z_2)| \leq \|z_1 - z_2\|_{\mathscr{V}}$. Pre-Gaussian uniformly in ν^r means that

$$\sup_r \mathbb{E}^r \Big[\sup_{f \in \mathscr{V}} \langle f, \xi^r \rangle \Big] < \infty. \tag{D.9}$$

Define $h_y : l^\infty(\mathscr{V}) \to \mathbb{R}$ by

$$h_y(\cdot) = \Big(\sup_{f\in\mathscr{V}} \langle f, \cdot\rangle - y + 1\Big)^+ \wedge 1.$$

Then it is clear that $h_y \in \mathrm{BL}_1$, and

$$\sup_r \mathbb{P}^r\Big(\sup_{f\in\mathscr{V}} \langle f, \xi_k^r\rangle > y\Big) \le \sup_r \mathbb{E}^r[h_y(\xi_k^r)].$$

By (D.8) and the above inequality, there exists $k_0 \in \mathbb{N}$ such that $k \ge k_0$ implies

$$\sup_r \mathbb{P}^r\Big(\sup_{f\in\mathscr{V}} \langle f, \xi_k^r\rangle > y\Big) \le \sup_r \mathbb{E}^r[h_y(\xi^r)] + y^{-q}.$$

Applying the definition of h_y and Markov inequality to obtain

$$\begin{aligned}\sup_r \mathbb{P}^r\Big(\sup_{f\in\mathscr{V}} \langle f, \xi_k^r\rangle > y\Big) &\le \sup_r \mathbb{P}^r\Big(\sup_{f\in\mathscr{V}} \langle f, \xi^r\rangle > y-1\Big) + y^{-q}\\ &\le y^{-q}\Big(2^q \sup_r \mathbb{E}^r\Big[\sup_{f\in\mathscr{V}} \langle f, \xi^r\rangle\Big]^q + 1\Big).\end{aligned}$$

Let M_q be the last term in parentheses, which does not depend on y. For each $r \in \mathbb{R}_+$, the Brownian bridge is separable and Gaussian with $\sup_{f\in\mathscr{V}} \langle f, \xi^r\rangle$ finite almost surely. Thus, there exist a constant M such that for all $r \in \mathbb{R}_+$,

$$\mathbb{E}^r\Big[\sup_{f\in\mathscr{V}} \langle f, \xi^r\rangle\Big]^q \le M\Big[\mathbb{E}^r \sup_{f\in\mathscr{V}} \langle f, \xi^r\rangle\Big]^q,$$

see Proposition A.2.4 in [25]. Conclude from (D.9) that $M_q < \infty$.

So far, we have shown that the result (D.7) is true for $n = 0$. Note that for any $n \in \mathbb{Z}$, $\xi_{n,k}^r$ is defined on the shifted sequence $v_{n+1}^r, v_{n+2}^r, \ldots$. By the i.i.d. property of the sequence, if we fix k then $\xi_{n,k}^r$ has the same distribution for all $n \in \mathbb{Z}$. So we can conclude that (D.7) is true for all $n \in \mathbb{Z}$. $\square$

Proof of Lemma D.1. Note that

$$|\langle f, \bar\eta^r(n,l)\rangle - l\langle f, \nu^r\rangle| \le \frac{1}{r}\sum_{i=n+1}^{n+\lfloor rl\rfloor} [\langle f, \delta_{v_i^r}\rangle - \langle f, \nu^r\rangle] + \frac{1}{r}.$$

Since for each $\varepsilon' > 0$, $1/r < \varepsilon'/2$ for all large r, so the probability in (D.4) can be bounded by

$$\limsup_{r\to\infty} \mathbb{P}^r\Bigg(\max_{-rM_1<n<r^2M_1} \sup_{l\in[0,L]} \sup_{f\in\mathscr{V}} \Bigg|\frac{1}{r}\sum_{i=n+1}^{n+\lfloor rl\rfloor} [\langle f, \delta_{v_i^r}\rangle - \langle f, \nu^r\rangle]\Bigg| > \frac{\varepsilon'}{2}\Bigg). \tag{D.10}$$

Pick $\delta > 0$, when r is large enough ($r > M_1/\delta$) the interval $[-rM_1, r^2M_1]$ will be covered by intervals

$$[-r^2\delta, 0], [0, r^2\delta], \ldots, \left[\left(\left\lceil \frac{M_1}{\delta} \right\rceil - 1\right) r^2\delta, \left\lceil \frac{M_1}{\delta} \right\rceil r^2\delta\right].$$

When r is large enough ($r^2\delta > \lfloor rL_1 \rfloor$), (D.10) can be further bounded by

$$\limsup_{r\to\infty} \mathbb{P}^r\left(\max_{-1\le j\le \lceil M_1/\delta\rceil -1} \max_{0\le k,k'\le r^2\delta} \sup_{f\in\mathscr{V}} \left| \frac{\sqrt{k}}{r}\langle f, \xi^r_{jr^2\delta,k}\rangle - \frac{\sqrt{k'}}{r}\langle f, \xi^r_{jr^2\delta,k'}\rangle \right| > \frac{\varepsilon'}{2}\right).$$

Since $\xi^r_{n,\cdot}$ has stationary increments, the previous term can be bounded above by

$$\limsup_{r\to\infty} \left\lceil \frac{M_1}{\delta} \right\rceil \mathbb{P}^r\left(\max_{0\le k\le r^2\delta} \sup_{f\in\mathscr{V}} \left| \frac{\sqrt{k}}{r}\langle f, \xi^r_{0,k}\rangle \right| > \frac{\varepsilon'}{2}\right).$$

By Ottaviani's inequality (see Proposition A.1.1 in [25]) and by stationary increments of $\xi^r_{n,\cdot}$, this can be bounded above by

$$\limsup_{r\to\infty} \frac{\lceil M_1/\delta\rceil \mathbb{P}^r(\sup_{f\in\mathscr{V}}\langle f, \xi^r_{0,\lfloor r^2\delta\rfloor}\rangle > \varepsilon'/(4\sqrt{\delta}))}{1 - \max_{0\le k\le r^2\delta} \mathbb{P}^r(\sup_{f\in\mathscr{V}}\langle f, \xi^r_{0,k}\rangle > \varepsilon' r/(4\sqrt{k}))}. \tag{D.11}$$

Assume δ is small enough so that $\frac{\varepsilon'}{4\sqrt{\delta}} > 2$. By Lemma D.2, there exists M_3 and $k_0 \in \mathbb{N}$ such that $k > k_0$ implies

$$\sup_r \mathbb{P}^r\left(\sup_{f\in\mathscr{V}}\langle f, \xi^r_{0,k}\rangle > \frac{\varepsilon'}{4\sqrt{\delta}}\right) \le \left(\frac{4\sqrt{\delta}}{\varepsilon'}\right)^3 M_3.$$

Since $\lfloor r^2\delta\rfloor \to \infty$ as $r \to \infty$, the limit superior of the numerator in (D.11) can be bounded above by $\lceil M_1/\delta\rceil (4\sqrt{\delta}/\varepsilon')^3 M_3$, which can be made arbitrarily small by choosing δ sufficiently small. By the same reason, those terms in the maximum of the denominator with index $k > k_0$ are bounded above by $(4\sqrt{\delta}/\varepsilon')^3 M_3$. For those terms with index $k \le k_0$,

$$\mathbb{P}^r\left(\sup_{f\in\mathscr{V}}\langle f, \xi^r_{0,k}\rangle > \frac{\varepsilon' r}{4\sqrt{k}}\right) \le \mathbb{P}^r\left(\sup_{f\in\mathscr{V}}\langle f, \xi^r_{0,k}\rangle > \frac{\varepsilon' r}{4\sqrt{k_0}}\right),$$

which converges to zero as $r \to \infty$. By choosing δ small enough, (D.11) can be made arbitrarily small for all large r. $\square$

Acknowledgments. The authors thank two anonymous referees for significantly improving the paper.

REFERENCES

[1] Asmussen, S. (2003). *Applied Probability and Queues: Stochastic Modelling and Applied Probability*, 2nd ed. *Applications of Mathematics (New York)* **51**. Springer, New York. MR1978607

[2] Avi-Itzhak, B. and Halfin, S. (1988). Expected response times in a non-symmetric time sharing queue with a limited number of service positions. In *Proceedings of the 12th International Teletraffic Congress*. Torino.

[3] Billingsley, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York. MR1324786

[4] Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed. Wiley, New York. MR1700749

[5] Blake, R. (1982). Optimal control of thrashing. In *Proceedings of the 1982 ACM SIGMETRICS Conference on Measurements and Modeling of Computer Systems*. Seattle, WA.

[6] Bramson, M. (1998). State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Syst.* **30** 89–148. MR1663763

[7] Carter, M. and van Brunt, B. (2000). *The Lebesgue–Stieltjes Integral: A Practical Introduction*. Springer, New York. MR1759133

[8] Denning, P. J., Kahn, K. C., Leroudier, J., Potier, D. and Suri, R. (1976). Optimal multiprogramming. *Acta Inform.* **7** 197–216. MR0449627

[9] Doytchinov, B., Lehoczky, J. and Shreve, S. (2001). Real-time queues in heavy traffic with earliest-deadline-first queue discipline. *Ann. Appl. Probab.* **11** 332–378. MR1843049

[10] Elnikety, S., Nahum, E., Tracy, J. and Zwaenepoel, W. (2004). A method for transparent admission control and request scheduling in e-commerce web sites. In *World Wide Web Conference*. New York.

[11] Ethier, S. N. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. Wiley, New York. MR0838085

[12] Gromoll, H. C. (2004). Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab.* **14** 555–611. MR2052895

[13] Gromoll, H. C. and Kruk, Ł. (2007). Heavy traffic limit for a processor sharing queue with soft deadlines. *Ann. Appl. Probab.* **17** 1049–1101. MR2326240

[14] Gromoll, H. C., Puha, A. L. and Williams, R. J. (2002). The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab.* **12** 797–859. MR1925442

[15] Gromoll, H. C., Robert, P. and Zwart, B. (2008). Fluid limits for processor-sharing queues with impatience. *Math. Oper. Res.* **33** 375–402. MR2415999

[16] Heiss, H. U. and Wagner, R. (1991). Adaptive load control in transaction processing systems. In *Proceedings of the 17th International Conference on Large Data Bases*. Barcelona.

[17] Kallenberg, O. (1986). *Random Measures*, 4th ed. Akademie Verlag, Berlin. MR0854102

[18] Kamra, A., Misra, V. and Nahum, E. M. (2004). Yaksha: A self-tuning controller for managing the performance of 3-tiered web sites. In *Twelfth IEEE International Workshop on Quality of Service*.

[19] Kaspi, H. and Ramanan, K. (2011). Law of large number limits for many-server queues. *Ann. Appl. Probab.* To appear.

[20] Kleinrock, L. (1976). *Queueing Systems. Volume 2: Computer Applications*. Wiley, New York.

[21] Nuyens, M. and van der Weij, W. (2007). The limited processor sharing queue. Technical report, CWI, Amsterdam.

[22] PUHA, A. L. and WILLIAMS, R. J. (2004). Invariant states and rates of convergence for a critical fluid model of a processor sharing queue. *Ann. Appl. Probab.* **14** 517–554. MR2052894

[23] RITCHIE, D. M. and THOMPSON, K. (1974). The Unix time-sharing system. *J. ACM* **17** 365–375.

[24] SCHROEDER, B., HARCHOL-BALTER, M., IYENGAR, A., NAHUM, E. and WIERMAN, A. (2006). How to determine a good multi-programming level for external scheduling. In *Proceedings of the 22nd International Conference on Data Engineering*. Atlanta, GA.

[25] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

[26] WILLIAMS, R. J. (1998). Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. *Queueing Syst.* **30** 27–88. MR1663759

[27] ZHANG, F. and LIPSKY, L. (2006). Modelling restricted processor sharing. In *Proc. of the 2006 Int'l Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA06)*.

[28] ZHANG, F. and LIPSKY, L. (2007). An analytical model for computer systems with non-exponential service times and memory thrashing overhead. In *Proc. of the 2007 Int'l Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA07)*.

[29] ZHANG, J., DAI, J. G. and ZWART, B. (2009). Law of large number limits of limited processor-sharing queues. *Math. Oper. Res.* **34** 937–970. MR2573503

J. ZHANG
DEPARTMENT OF INDUSTRIAL ENGINEERING
AND LOGISTIC MANAGEMENT
HONG KONG UNIVERSITY OF SCIENCE
AND TECHNOLOGY
CLEAR WATER BAY
HONG KONG
E-MAIL: jiheng@ust.hk

J. G. DAI
H. MILTON STEWART SCHOOL
OF INDUSTRIAL AND SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332
USA
E-MAIL: dai@gatech.edu

B. ZWART
CWI
P.O. BOX 94079
1090 GB AMSTERDAM
THE NETHERLANDS
E-MAIL: bert.zwart@cwi.nl