Operations Research Letters 42 (2014) 388–393

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Convergence to equilibrium states for fluid models of many-server queues with abandonment

Zhenghua Long, Jiheng Zhang [*]

Department of Industrial Engineering and Logistic Management, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 2 June 2014
Received in revised form
20 June 2014
Accepted 24 June 2014
Available online 1 July 2014

Keywords:
Many-server queue
Abandonment
Fluid model
Equilibrium state
Convergence

ABSTRACT

Fluid models, in particular their equilibrium states, have become an important tool for the study of many-server queues with general service and patience time distributions. However, it remains an open question whether the solution to a fluid model converges to the equilibrium state and under what condition. We show in this paper that the convergence holds under some conditions. Our method builds on the framework of measure-valued processes, which keeps track of the remaining patience and service times.

1. Introduction

In this paper, we analyze the asymptotic behavior of fluid models for many-server queues with abandonment. We allow both the service time and patience time distributions to be general. To the best of our knowledge, Whitt [10] is the earliest to propose a fluid model for many-server queues with generally distributed service and patience times. In [10], the equilibrium state for a fluid model is characterized and extensive simulations show that the equilibrium state of the fluid model yields reasonably good approximations to the original stochastic system in steady state.

The challenge in studying many-server queues, especially when the service time is generally distributed, is that the status of the server pool plays an important role in the dynamics. However, describing the status itself is quite complicated. There have been two streams of work providing different modeling approaches. Kang and Ramanan [5], which is based on [6] for many-server queues without abandonment, modeled the status of the server pool by keeping track of the "age" (the amount of time a customer has been in service). Alternatively, Zhang [11] modeled the status of the server pool by tracking each customer's "residual" (the remaining service time). The fluid model proposed in [5] is too complicated to be analyzed. Even the existence and uniqueness of the fluid model solution is proved using heavy traffic approximation. This paper thus builds on the second approach instead.

Both [5,11] established the fluid model as the limit of fluid-scaled stochastic processes underlying many-server queues. However, the analysis of the fluid model itself remains open. [10,5,11] have all been unable to show that the fluid model converges to the equilibrium states. Such a convergence was proved in [9] for a many-server fluid model with exponentially distributed service and patience times. Taking advantage of the exponential distribution, the fluid model reduces to a one-dimensional ordinary differential equation (ODE). In general, proving convergence to the equilibrium states for fluid models is intrinsically difficult, even though the fluid models are just deterministic dynamic systems.

The current work can be viewed as a sequel to [11]. We use the same definition for the fluid model, and even the same set of notations for easy connection. The modeling is close to that in [12] but the method is significantly different due to customer abandonment (which does not appear in [12]) and intrinsic difficulties in many-server models. [7] offered a nice treatment for the fluid model of the many-server queue without abandonment. Though the main focus of that paper is not the fluid analysis, the elegant treatment of the fluid model helps to relax the assumption on initial customers made in [8]. Abandonment, especially with a general patience time distribution, imposes significant challenges. A virtual buffer, which holds all the customers who have arrived but not yet scheduled to receive service according to the FCFS policy, is constructed to study abandonment in [11]. The idea is to keep some abandoned customers in the virtual buffer for tracking purposes. This paper

[*] Correspondence to: Clear Water Bay, Hong Kong Special Administrative Region.
E-mail address: jiheng@ust.hk (J. Zhang).

http://dx.doi.org/10.1016/j.orl.2014.06.011
0167-6377/© 2014 Published by Elsevier B.V.

adopts the same idea. Our fluid model can be shown to be equivalent to the one in [7] when patience time becomes infinite (no abandonment).

We hope the analytical tools we develop in this paper can pave the way for studying more complicated many-server models such as the multi-class V-model studied in [1], and models where service and patience times are dependent in [2].

2. Fluid models of many-server queues

Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{R}_+ = [0, \infty)$. For $a, b \in \mathbb{R}$, write a^+ for the positive part of a and $a \wedge b$ for the minimum. Denote $C_x = (x, \infty)$ and $F^c(x) = 1 - F(x)$ for any distribution function. At time t, let $\bar{\mathscr{R}}(t)(C_x)$ denote the amount of fluid in the virtual buffer with remaining patience time larger than x. Since the virtual buffer also holds abandoned customers who have negative remaining patience times, the testing parameter x is allowed to be both positive and negative for the measure $\bar{\mathscr{R}}(t)$. Introduce $\bar{R}(t) = \bar{\mathscr{R}}(t)(\mathbb{R})$, the total fluid content in the virtual buffer. Denote by λ the arrival rate. So at time t, the earliest arrived fluid content in the virtual buffer arrives at time $t - \bar{R}(t)/\lambda$. To find out the status of the virtual buffer at time t, we take integral from $t - \bar{R}(t)/\lambda$ to t. If an infinitesimal amount of fluid content λds arrives at time s, only a fraction $F^c(x+t-s)$ of it has remaining patience time larger than x at time t since $t - s$ amount of time has been spent waiting in queue. This yields Eq. (2.2). Let $\bar{\mathscr{Z}}(t)(C_x)$ denote the amount of fluid in the server pool with remaining service time larger than x at time t. Unlike the virtual buffer, a customer leaves the system once his remaining service time hits 0. So we restrict the testing parameter $x \in \mathbb{R}_+$ for the measure $\bar{\mathscr{Z}}(t)$. Let

$$\bar{B}(t) = \lambda t - \bar{R}(t). \tag{2.1}$$

The physical intuition for $\bar{B}$ is that $\bar{B}(t) - \bar{B}(s)$ represents the amount of fluid in the virtual buffer that could have entered service during time interval $(s, t]$. It should be pointed out that not all of it will actually enter the server pool. At time s, an infinitesimal amount $d\bar{B}(s)$ is scheduled to enter service after waiting in the virtual buffer for $\bar{R}(s)/\lambda$. Thus, a fraction $F\left(\frac{\bar{R}(s)}{\lambda}\right)$ has actually abandoned queue by time s. Only the rest makes it to the service. This contributes to the term $F^c\left(\frac{\bar{R}(s)}{\lambda}\right)$ in (2.3). The following *fluid dynamic equations* characterize how the fluid content $(\bar{\mathscr{R}}(t), \bar{\mathscr{Z}}(t))$ evolves over time. For all $t \geq 0$,

$$\bar{\mathscr{R}}(t)(C_x) = \lambda \int_{t-\frac{\bar{R}(t)}{\lambda}}^{t} F^c(x+t-s)ds, \quad x \in \mathbb{R}, \tag{2.2}$$

$$\bar{\mathscr{Z}}(t)(C_x) = \bar{\mathscr{Z}}(0)(C_{x+t}) + \int_0^t F^c\left(\frac{\bar{R}(s)}{\lambda}\right) G^c(x+t-s)d\bar{B}(s),$$
$$x \in \mathbb{R}_+. \tag{2.3}$$

Introduce $\bar{Z}(t) = \bar{\mathscr{Z}}(t)(C_0)$, the fluid content in service; and $\bar{Q}(t) = \bar{\mathscr{R}}(t)(C_0)$, the fluid queue length. Let $\bar{Z}(t) + \bar{Q}(t) = \bar{X}(t)$ denote the total amount of fluid in the physical system. The following non-idling constraints must be valid at any time $t \geq 0$,

$$\bar{Q}(t) = (\bar{X}(t) - 1)^+, \tag{2.4}$$

$$\bar{Z}(t) = \bar{X}(t) \wedge 1. \tag{2.5}$$

Let (λ, F, G) denote the *fluid model* defined by (2.2)–(2.5). The initial state $(\bar{\mathscr{R}}(0), \bar{\mathscr{Z}}(0))$ is said to be *valid* if it satisfies Eqs. (2.2)–(2.5) at time $t = 0$. Throughout this paper, we make the following assumptions.

Assumption 1. Assume the service time distribution G is absolutely continuous with finite mean $1/\mu$; and the patience time distribution F is Lipschitz continuous.

According to Theorem 3.1 in [11], under Assumption 1, there exists a unique solution to the fluid model (λ, F, G) for any valid initial state $(\bar{\mathscr{R}}(0), \bar{\mathscr{Z}}(0))$. Theorem 3.3 in [11] shows that the fluid model solution serves as the fluid limit of the many-server queueing models.

3. Convergence to equilibrium states

A key property of the fluid model is that it has an equilibrium state. An equilibrium state is defined intuitively as the state from which the fluid model solution starts and remains. More precisely, $(\bar{\mathscr{R}}_\infty, \bar{\mathscr{Z}}_\infty)$ is an *equilibrium state* of the fluid model (λ, F, G) if the solution to the fluid model with a valid initial state $(\bar{\mathscr{R}}_\infty, \bar{\mathscr{Z}}_\infty)$ satisfies $(\bar{\mathscr{R}}(t), \bar{\mathscr{Z}}(t)) = (\bar{\mathscr{R}}_\infty, \bar{\mathscr{Z}}_\infty)$ for all $t \geq 0$. As characterized in Theorem 3.2 in [11], the state $(\bar{\mathscr{R}}_\infty, \bar{\mathscr{Z}}_\infty)$ is an equilibrium state of the fluid model (λ, F, G) if and only if it satisfies

$$\bar{\mathscr{R}}_\infty(C_x) = \lambda \int_0^\omega F^c(x+s)ds, \quad x \in \mathbb{R}, \tag{3.1}$$

$$\bar{\mathscr{Z}}_\infty(C_x) = \min(\rho, 1)\,[1 - G_e(x)], \quad x \in \mathbb{R}_+, \tag{3.2}$$

where $\rho = \lambda/\mu$ is the traffic intensity, ω is the *unique* solution to

$$F(\omega) = \max\left(\frac{\rho - 1}{\rho}, 0\right), \tag{3.3}$$

and $G_e(\cdot)$, called the equilibrium distribution associated with G, is defined by

$$G_e(x) = \mu \int_0^x G^c(y)dy, \quad \text{for all } x \geq 0. \tag{3.4}$$

Note that we need to assume (3.3) has a unique solution (see Remark 1 for detailed discussion). The objective is to show

$$\lim_{t \to \infty} \left(\bar{\mathscr{R}}(t), \bar{\mathscr{Z}}(t)\right) = \left(\bar{\mathscr{R}}_\infty, \bar{\mathscr{Z}}_\infty\right). \tag{3.5}$$

Underloaded case. In this case, we can prove the convergence under a fairly general condition. We only require the initial state to satisfy

$$\lim_{x \to \infty} \bar{\mathscr{Z}}(0)(C_x) = 0, \tag{3.6}$$

which is quite mild. We do not even require the initial remaining workload in the server pool $\int_0^\infty \bar{\mathscr{Z}}(0)(C_x)dx$ to be finite.

Theorem 1. *Under Assumption 1 and suppose $\lambda < \mu$, if the initial state satisfies (3.6), then the convergence (3.5) holds.*

Critically loaded and overloaded cases. The study in these two cases turns out to be more challenging. We cannot prove that the convergence holds in generality. If the initial state is controlled by (3.7), we can prove the convergence without assuming additional conditions on service and patience time distributions. This condition covers the cases where the system starts from empty or initial customers' service times follow the equilibrium distribution.

Theorem 2. *Under Assumption 1 and suppose $\lambda \geq \mu$, if there is a unique solution to (3.3), the initial state satisfies (3.6) and*

$$\bar{\mathscr{Z}}(0)'((0, t]) := \frac{d}{dt}\bar{\mathscr{Z}}(0)((0, t]) \leq \lambda G^c(t), \tag{3.7}$$

then the convergence (3.5) holds.

4. Preliminary analysis

Introduce two new functions $F_d(x) = \int_0^x F^c(y)dy$ and

$$H(x) = \begin{cases} F^c\left(F_d^{-1}\left(\dfrac{x}{\lambda}\right)\right), & \text{if } 0 \leq x < \lambda N_F, \\ 0, & \text{if } x \geq \lambda N_F, \end{cases} \tag{4.1}$$

where N_F is the mean of the patience time, i.e., $N_F = \int_0^\infty F^c(y)dy$, which can be either finite or infinite. In [11], the following key equation is derived from the fluid equations (2.2)–(2.5) and will play an import role in this paper,

$$\begin{aligned}\bar X(t) &= \bar{\mathcal{Z}}(0)(C_t) + \bar Q(0)G^c(t)\\ &\quad + \frac{\lambda}{\mu}\int_0^t H\left((\bar X(t-s)-1)^+\right) dG_e(s)\\ &\quad + \int_0^t (\bar X(t-s)-1)^+ dG(s). \end{aligned} \tag{4.2}$$

We introduce an auxiliary *enter service* process

$$\bar A(t) = \int_0^t F^c\left(\frac{\bar R(s)}{\lambda}\right) d\bar B(s).$$

According to the fluid dynamic equation (2.2),

$$\bar{\mathcal{R}}(t)(C_x) = \lambda \int_0^{\frac{\bar R(t)}{\lambda}} F^c(x+s)ds. \tag{4.3}$$

Plugging $x = 0$ into the above equation gives

$$\bar Q(t) = \lambda \int_{t-\frac{\bar R(t)}{\lambda}}^t F^c(t-s)ds = \lambda \int_0^{\frac{\bar R(t)}{\lambda}} F^c(s)ds. \tag{4.4}$$

Utilizing (2.1), (4.1) and (4.4), and the fact that $\bar Q(t)$ (equivalently $\bar R(t)$) is of bounded total variation (see p. 162 in [11]), the auxiliary process can be written as

$$\bar A(t) = \lambda \int_0^t H(\bar Q(s))ds - \bar Q(t) + \bar Q(0). \tag{4.5}$$

It follows from Lemma A.3 in [11] that $\bar A(t)$ is non-decreasing in t. Introduce the *abandonment* process

$$\bar L(t) = \lambda \int_0^t F\left(\frac{\bar R(s)}{\lambda}\right) ds = \lambda t - \lambda \int_0^t H(\bar Q(s))ds, \tag{4.6}$$

where the second equation can be verified from (4.1) and (4.4). This together with (4.5) implies the balance equation

$$\bar Q(t) = \bar Q(0) + \lambda t - \bar L(t) - \bar A(t). \tag{4.7}$$

According to the fluid dynamic equation (2.3),

$$\bar{\mathcal{Z}}(t)(C_x) = \bar{\mathcal{Z}}(0)(C_{x+t}) + \int_0^t G^c(x+t-s)d\bar A(s). \tag{4.8}$$

Plugging $x = 0$ to (4.8) and performing integration by parts yield the relation between $\bar A$ and $\bar Z$

$$\bar A(t) = \bar Z(t) - \bar{\mathcal{Z}}(0)(C_t) + \int_0^t \bar A(t-s)dG(s). \tag{4.9}$$

Let G^{n*} be the n-fold convolution of G with itself, and denote $M_G(t) = \sum_{i=1}^\infty G^{n*}(t)$ as the renewal function of G. The solution to the above renewal equation is

$$\bar A(t) = \left(\bar Z(t) - \bar{\mathcal{Z}}(0)(C_t)\right) * U_G(t), \tag{4.10}$$

where $U_G(t) = M_G(t) + 1$. We can also introduce the *service completion* process

$$\begin{aligned}\bar S(t) &= \bar{\mathcal{Z}}(0)((0,t]) + \int_0^t G(t-s)d\bar A(s)\\ &= \bar{\mathcal{Z}}(0)((0,t]) + (\bar Z(t) - \bar{\mathcal{Z}}(0)(C_t)) * M_G(t). \end{aligned} \tag{4.11}$$

By (4.10) and (4.11) one can verify the balance equation

$$\bar Z(t) = \bar Z(0) + \bar A(t) - \bar S(t). \tag{4.12}$$

5. Proof of the convergence

The proof is made possible by carefully analyzing the measure-valued fluid dynamic equations (2.2)–(2.3) and the above introduced auxiliary processes. The most important step is to show the convergence of total amount of the fluid process $\bar X(t)$. However, the measure-valued processes play a significant role in analyzing the real-valued process $\bar X(t)$.

Proposition 1 (*Underloaded*)**.** *Under the same conditions in Theorem 1, the fluid model solution $(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t))$ satisfies*

$$\lim_{t\to\infty} \bar X(t) = \frac{\lambda}{\mu}. \tag{5.1}$$

Proof. By (2.4) and (4.2) we have,

$$\begin{aligned}\bar Q(t) &= -\bar Z(t) + \bar{\mathcal{Z}}(0)(C_t) + \bar Q(0)G^c(t)\\ &\quad + \frac{\lambda}{\mu}\int_0^t H(\bar Q(t-s))dG_e(s) + \int_0^t \bar Q(t-s)dG(s). \end{aligned}$$

Define

$$\begin{aligned}\bar K(t) &= -\bar Z(t) + \bar{\mathcal{Z}}(0)(C_t) + \bar Q(0)G^c(t)\\ &\quad + \frac{\lambda}{\mu}\int_0^t H(\bar Q(t-s))dG_e(s), \end{aligned} \tag{5.2}$$

then

$$\bar Q(t) = \bar K(t) + \int_0^t \bar Q(t-s)dG(s). \tag{5.3}$$

So

$$\bar Q(t) = \bar K(t) * U_G(t) = \int_0^t \bar K(t-s)dU_G(s). \tag{5.4}$$

Consider the following two cases: If $\bar Q(t) = 0$, then by (5.3),

$$\begin{aligned}\bar K(t) &= \bar Q(t) - \int_0^t \bar Q(t-s)dG(s)\\ &= 0 - \int_0^t \bar Q(t-s)dG(s) \le 0. \end{aligned}$$

If $\bar Q(t) > 0$, then $\bar Z(t) = 1$ due to non-idling constraints. Since $\lambda < \mu$ and $H(\cdot) \le 1$, we can pick $\delta = (1 - \lambda/\mu)/3$ which is positive such that $\frac{\lambda}{\mu}\int_0^t H(\bar Q(t-s))dG_e(s) \le 1 - 2\delta$. For this given $\delta > 0$, there exists a T such that $\bar{\mathcal{Z}}(0)(C_t) + \bar Q(0)G^c(t) \le \delta$ for all $t \ge T$. It now follows from (5.2) that for all $t \ge T$ and $\bar Q(t) > 0$,

$$\begin{aligned}\bar K(t) &= -1 + \bar{\mathcal{Z}}(0)(C_t) + \bar Q(0)G^c(t) + \frac{\lambda}{\mu}\int_0^t H(\bar Q(t-s))dG_e(s)\\ &\le -1 + \delta + 1 - 2\delta = -\delta. \end{aligned}$$

Denote by the set $\mathcal{S} = \{t \ge 0 : \bar Q(t) > 0\}$ the collection of times when the fluid queue is positive. We first prove by contradiction that $m(\mathcal{S}) < \infty$, where m is the Lebesgue measure on real numbers. Now suppose $m(\mathcal{S}) = \infty$. Combining the above two cases, we have $\bar K(t) \le 0$ for all $t \in [T, +\infty)$ and $\bar K(t) \le -\delta$ for all $t \in \mathcal{S} \cap [T, +\infty)$.

Write the integral in (5.4) in two parts,

$$\bar Q(t) = \int_0^{t-T} \bar K(t-s)dU_G(s) + \int_{t-T}^t \bar K(t-s)dU_G(s). \tag{5.5}$$

It is clear that the first term in (5.5) is negative whenever $t > T$ since $\bar K(t-s) \le 0$ for all $s \in [0, t-T]$. Moreover $\bar K(t-s) \le -\delta$

for all $s \in \mathcal{S}_{t,T}$, where $\mathcal{S}_{t,T} := \{s : t - s \in \mathcal{S} \cap [T, \infty)\}$. By the assumption that $m(\mathcal{S}) = \infty$, we have $m(\mathcal{S}_{t,T}) \to \infty$ as $t \to \infty$. So the first term on the right-hand side of (5.5)

$$\int_0^{t-T} \bar{K}(t-s)dU_G(s) \le \int_{\mathcal{S}_{t,T}} -\delta dU_G(s),$$

which converges to $-\infty$ as $t \to \infty$. The second term on the right-hand side of (5.5) is essentially an integral on a finite interval. Let $M = \sup_{s\in[0,T]} \bar{K}(s)$, then

$$\int_{t-T}^{t} \bar{K}(t-s)dU_G(s) \le M \int_{t-T}^{t} dU_G(s) \to \mu MT,$$

as $t \to \infty$. So we have $\lim_{t\to\infty} \bar{Q}(t) = -\infty$, which contradicts $\bar{Q}(t) \ge 0$ for all $t \ge 0$. Thus we have proved $m(\mathcal{S}) < \infty$. As a byproduct, the above analysis also yields an upper bound for the fluid queue length process. Since the first term on the right-hand side of (5.5) is always less than or equal to 0 for all $t \ge T$, and the second term has an asymptotic upper bound, there exists a constant M_1 such that

$$\sup_{t\ge 0} \bar{Q}(t) \le M_1 + \mu MT. \tag{5.6}$$

Since $m(\mathcal{S}) < \infty$, for any $\varepsilon > 0$ there exists a τ such that $m(\mathcal{S} \cap [\tau, +\infty)) < \varepsilon$. Now consider the fluid model shifted by time τ as in Lemma 3. Let $S_{\tau+t} := \{s : \tau + t - s \in \mathcal{S} \cap [\tau, \infty)\}$, then

$$m\left(\mathcal{S}_{\tau+t}\right) \le m\left(\mathcal{S} \cap [\tau, +\infty)\right) < \varepsilon. \tag{5.7}$$

Since $G(\cdot)$ is absolutely continuous by Assumption 1, we can choose an ε small enough and a corresponding τ such that

$$\int_0^t \bar{Q}_\tau(t-s)dG(s) \le (M_1 + \mu MT) \int_{\mathcal{S}_{\tau+t}} dG(s) \le \frac{1}{2}\left(1 - \frac{\lambda}{\mu}\right),$$

where the first inequality is due to (5.6) and the second one follows from (5.7), Theorem 12.34 in [4] and the fact that $\lambda/\mu < 1$. Now by (A.4) in Lemma 3 and that $H(\cdot) \le 1$,

$$\bar{X}_\tau(t) \le \bar{\mathcal{Z}}(\tau)(C_t) + \bar{Q}(\tau)G^c(t) + \frac{\lambda}{\mu} + \frac{1}{2}\left(1 - \frac{\lambda}{\mu}\right).$$

Replacing (t, x) in (4.8) by (τ, t), it follows from the monotonicity of $\bar{A}(\cdot)$ and (3.6) that $\bar{\mathcal{Z}}(\tau)(C_t)$ vanishes as $t \to \infty$. Therefore there exists a τ_1 such that $\bar{X}(t) < 1$, consequently $\bar{Q}(t) = 0$, for all $t \ge \tau_1$. Then by (4.2), (2.4) and Lemma 2, (5.1) immediately follows. □

Proposition 2 (*Critically Loaded and Overloaded*). *Under the same conditions in Theorem 2, if (3.3) has a unique solution ω, then the fluid model solution $(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t))$ satisfies*

$$\lim_{t\to\infty} \bar{X}(t) = 1 + \lambda \int_0^\omega F^c(x)dx. \tag{5.8}$$

Proof. Let $T = \inf\{t \ge 0 : \bar{Z}(t) = 1\}$. So (2.4), (4.5) and (4.8) reveal for any $t \in [0, T]$ we have

$$\bar{Z}(t) = \bar{\mathcal{Z}}(0)(C_t) + \lambda \int_0^t G^c(s)ds, \tag{5.9}$$

and $\bar{Z}(T) = \bar{\mathcal{Z}}(0)(C_T) + \lambda \int_0^T G^c(s)ds = 1$. One can see $T = \infty$ only when $\lambda = \mu$. In this case, it is clear that $\lim_{t\to\infty} \bar{Z}(t) = 1$ and (5.8) holds. Thus, we focus on the case where $T < \infty$.

It is easy to verify that under Assumption 1 and (3.7), the fluid model solution becomes differentiable. Take derivative of (4.11) to obtain

$$\bar{S}'(t) = \bar{\mathcal{Z}}(0)'((0,t]) + \int_0^t M'_G(t-s)d(\bar{Z}(s) - \bar{\mathcal{Z}}(0)(C_s)). \tag{5.10}$$

For any $t \in [0, T]$, plug in (5.9) and apply condition (3.7)

$$\begin{aligned}\bar{S}'(t) &= \bar{\mathcal{Z}}(0)'((0,t]) + \lambda \int_0^t M'_G(t-s)G^c(s)ds \\ &\le \lambda G^c(t) + \lambda G(t) = \lambda;\end{aligned}$$

for any $t \in (T, \infty)$, split the integral in (5.10) and apply (3.7)

$$\begin{aligned}\bar{S}'(t) &= \bar{\mathcal{Z}}(0)'((0,t]) + \lambda \int_0^T M'_G(t-s)G^c(s)ds \\ &\quad - \int_T^t M'_G(t-s)d\bar{\mathcal{Z}}(0)(C_s) + \int_T^t M'_G(t-s)d\bar{Z}(s) \\ &\le \lambda + \int_T^t M'_G(t-s)d\bar{Z}(s).\end{aligned} \tag{5.11}$$

Specializing the patience time distribution to be $F(x) = 0$ for all $x \ge 0$ yields the fluid model with infinite patience (no-abandonment). This implies the remaining patience time of all the fluid in (virtual) buffer is $+\infty$. So we need to extend the real line to include $+\infty$. Let $\bar{\mathcal{R}}_p$, $\bar{\mathcal{Z}}_p$, $\bar{R}_p$, $\bar{Q}_p$, $\bar{Z}_p$, $\bar{X}_p$, $\bar{A}_p$, $\bar{L}_p$ and $\bar{S}_p$ denote the associated processes of the fluid model with infinite patience. We replace C_x by $C^*_x = C_x \cup \{+\infty\}$ when discussing the measure $\bar{\mathcal{R}}_p$ for the virtual buffer, e.g., $\bar{Q}_p(t) = \bar{\mathcal{R}}_p(t)(C^*_0)$. It is clear that the measure-valued process $(\bar{\mathcal{R}}_p(t), \bar{\mathcal{Z}}_p(t))$ still satisfies (2.2)–(2.3) with the constraints (2.4)–(2.5) remaining the same. All we need is to plug in $F(\cdot) \equiv 0$ to obtain the corresponding version for the no-abandonment model. E.g., the terms $F^c(\cdot)$ in (2.2) and (2.3) become 1. So $\bar{\mathcal{R}}_p(t)(C^*_x) = \bar{R}_p(t) = \bar{Q}_p(t)$ by (2.2). To the other extreme, specializing the patience time distribution to be $F(x) = 1$ for all $x \ge 0$ yields the blocked model (no buffer). But this is not simply an extension of our fluid model. So we define it in Definition 1. Denote by $\bar{Q}_b$, $\bar{Z}_b$ and $\bar{A}_b$ the processes associated with the blocked fluid model. For these two fluid models, we can explicitly solve them (see (4.5), (4.10), (5.9) and (5.11))

$$\bar{Z}_p(t) = \begin{cases} \bar{\mathcal{Z}}(0)(C_t) + \lambda \displaystyle\int_0^t G^c(s)ds, & t \in [0, T], \\ 1, & t \in (T, \infty), \end{cases} \tag{5.12}$$

$$\bar{Q}_p(t) = \bar{Q}(0) + \lambda t - (\bar{Z}_p(t) - \bar{\mathcal{Z}}(0)(C_t)) * U_G(t),$$

and $\bar{Z}_b(t) = \bar{Z}_p(t)$, $\bar{Q}_b(t) = 0$. Consequently $\bar{A}_b(t) = \bar{A}_p(t)$ due to (4.10). It follows from Lemma 1 and Corollary 1 that $\bar{A}_b(t) = \bar{A}(t) = \bar{A}_p(t)$. So we can conclude that $\bar{Z}$ also satisfies (5.12) due to (4.9). Combining this result with assumption (3.7), it is easily seen that $\bar{Z}'(t) - \bar{\mathcal{Z}}(0)'(C_t)$ is directly Riemann integrable. Applying the key renewal theorem to the differentiated version of (4.9) yields that

$$\lim_{t\to\infty} \bar{A}'(t) = \mu \int_0^\infty \bar{Z}'(t) - \bar{\mathcal{Z}}(0)'(C_t)dt = \mu. \tag{5.13}$$

Next, we prove the convergence of $\bar{Q}(t)$ in two cases. Case 1: $\lambda = \mu$. Since we assume there is a unique solution to (3.3) (the solution $\omega = 0$ in this case), it then follows from the definition of $H(\cdot)$ in (4.1) that for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $\lambda H(\bar{Q}(t)) \le \mu - \delta$ whenever $\bar{Q}(t) \ge \varepsilon$. Due to (5.13), there exists a $T_0 > 0$ such that for all $t > T_0$, $\bar{A}'(t) \ge \mu - \delta/2$. Let $\mathcal{L}(t) = (\bar{Q}(t) - 0)^2$, then by (4.5) for all $t > T_0$

$$\mathcal{L}'(t) = 2(\bar{Q}(t) - 0)(\lambda H(\bar{Q}(t)) - \bar{A}'(t)) \le -\varepsilon\delta,$$

whenever $\bar{Q}(t) \ge \varepsilon$. So there must be a $T_1 > 0$ such that $\bar{Q}(t) < \varepsilon$ for any $t > T_1$. Due to the arbitrariness of ε, we have

$\lim_{t\to\infty} \bar{Q}(t) = 0$. Case 2: $\lambda > \mu$. Let $\bar{Q}_\infty = \lambda \int_0^\omega F^c(x)dx$. Since ω is the unique solution to (3.3), similar to the previous case, for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\lambda H(\bar{Q}(t)) \le \mu - \delta \quad \text{whenever } \bar{Q}(t) \ge \bar{Q}_\infty + \varepsilon, \tag{5.14}$$

$$\lambda H(\bar{Q}(t)) \ge \mu + \delta \quad \text{whenever } \bar{Q}(t) \le \bar{Q}_\infty - \varepsilon. \tag{5.15}$$

Due to (5.13), there exists a $T_0 > 0$ such that for all $t > T_0$, $\mu - \delta/2 \le \bar{A}'(t) \le \mu + \delta/2$. Let $\mathcal{L}(t) = (\bar{Q}(t) - \bar{Q}_\infty)^2$, then by (4.5) for all $t > T_0$

$$\mathcal{L}'(t) = 2(\bar{Q}(t) - \bar{Q}_\infty)(\lambda H(\bar{Q}(t)) - \bar{A}'(t)) \le -\varepsilon\delta,$$

whenever $\bar{Q}(t) \le \bar{Q}_\infty - \varepsilon$ or $\bar{Q}(t) \ge \bar{Q}_\infty + \varepsilon$. So there must be a $T_1 > 0$ such that $\bar{Q}(t) \in (\bar{Q}_\infty - \varepsilon, \bar{Q}_\infty + \varepsilon)$ for all $t > T_1$. Due to the arbitrariness of ε, we have $\lim_{t\to\infty} \bar{Q}(t) = \bar{Q}_\infty$. □

Remark 1. In the critically loaded and overloaded cases, if the solution to (3.3) is not unique then define

$$\bar{Q}_{\infty,\max} := \sup\left\{\lambda \int_0^\omega F^c(x)dx : F(\omega) = \frac{\rho - 1}{\rho}\right\}.$$

We can use (5.14) to show $\limsup_{t\to\infty} \bar{Q}(t) \le \bar{Q}_{\infty,\max}$. Similarly, we can define $\bar{Q}_{\infty,\min}$ and use (5.15) to show $\liminf_{t\to\infty} \bar{Q}(t) \ge \bar{Q}_{\infty,\min}$.

Proving the convergence of the measure-valued process from that of $X(t)$ is the same for systems with different load.

Complete the proof of Theorems 1–2. Since the space of real numbers is separable and $\bar{\mathcal{R}}_\infty(\{x\}) = \bar{\mathcal{Z}}_\infty(\{x\}) = 0$ for all x, according to Property (iv) of the Prokhorov metric on p. 72 in [3], it suffices to show that

$$\lim_{t\to\infty} \bar{\mathcal{R}}(t)(C_x) = \bar{\mathcal{R}}_\infty(C_x), \tag{5.16}$$

$$\lim_{t\to\infty} \bar{\mathcal{Z}}(t)(C_x) = \bar{\mathcal{Z}}_\infty(C_x). \tag{5.17}$$

It follows from Propositions 1–2 that the limit of $\bar{Q}(t)$ is

$$\lim_{t\to\infty} \bar{Q}(t) = \bar{Q}_\infty = \lambda \int_0^\omega F^c(x)dx, \tag{5.18}$$

where ω satisfies (3.3). Then by (4.1) and (4.4), $\lim_{t\to\infty} \bar{R}(t) = \bar{R}_\infty = \lambda\omega$ and

$$\lim_{t\to\infty} H(\bar{Q}(t)) = H(\bar{Q}_\infty) = \begin{cases} 1, & \lambda \le \mu, \\ \mu/\lambda, & \lambda > \mu. \end{cases}$$

Hence, (5.16) follows from (3.1), (4.3) and the above limit. Plugging (4.5) into (4.8) yields

$$\begin{aligned} \bar{\mathcal{Z}}(t)(C_x) &= \bar{\mathcal{Z}}(0)(C_{x+t}) + \lambda \int_0^t H(\bar{Q}(t-s))G^c(x+s)ds \\ &\quad - \bar{Q}(t)G^c(x) + \bar{Q}(0)G^c(x+t) \\ &\quad - \int_0^t \bar{Q}(t-s)dG^c(x+s). \end{aligned}$$

According to Lemma 2 and the convergence of $\bar{Q}(t)$ in (5.18),

$$\lim_{t\to\infty} \bar{\mathcal{Z}}(t)(C_x) = \frac{\lambda H(\bar{Q}_\infty)}{\mu}\left(1 - \mu \int_0^x G^c(s)ds\right).$$

So (5.17) follows from (3.2) and (3.4). □

Acknowledgment

The research is supported in part by the Hong Kong Research Grants Council (Grants GRF-622411 and GRF-624012).

Appendix. Some auxiliary lemmas

A comparison result. Consider two fluid models with the same arrival rate λ, service time distribution G, and initial state, but different patience time distributions F_i, $i = 1, 2$. Denote by $\bar{Q}_i$, $\bar{Z}_i$, $\bar{A}_i$ and $\bar{L}_i$ the corresponding derived processes associated with the ith fluid model.

Lemma 1. *If $F_1^c(x) \le F_2^c(x)$ for all $x \in \mathbb{R}_+$, then $\bar{A}_1(t) \le \bar{A}_2(t)$ for all $t \ge 0$.*

Proof. For any $\delta > 0$, let $\tau = \inf\{t \in \mathbb{R}_+ : \bar{A}_1(t) - \bar{A}_2(t) \ge \delta\}$ be the first time when $\bar{A}_1$ exceeds $\bar{A}_2$ by δ. Since the two fluid models start from the same initial state, we must have $\tau > 0$. Now the objective is to show that $\tau = \infty$. Suppose τ is finite.

For any $t \in [0, \tau]$, if $\bar{Z}_1(t) \le \bar{Z}_2(t)$, then by (4.9)

$$\begin{aligned} \bar{A}_1(t) - \bar{A}_2(t) &= \bar{Z}_1(t) - \bar{Z}_2(t) - \int_0^t (\bar{A}_1(s) - \bar{A}_2(s))dG(t-s) \\ &\le - \int_0^t (\bar{A}_1(s) - \bar{A}_2(s))dG(t-s) \\ &< \delta G(t) \le \delta, \end{aligned} \tag{A.1}$$

for any $t \in [0, \tau]$. This implies that $\bar{Z}_1(\tau) > \bar{Z}_2(\tau)$. A direct consequence is that

$$\bar{Q}_2(\tau) = 0, \tag{A.2}$$

due to the non-idling equations (2.4) and (2.5). Let $r = \sup\{t < \tau : \bar{Q}_1(t) < \bar{Q}_2(t)\} \vee 0$ be the last time $\bar{Q}_1$ is less than $\bar{Q}_2$. Thus $\bar{Q}_1(t) \ge \bar{Q}_2(t)$ for each $t \in [r, \tau]$. Then it follows from (4.4) and the fact $F_1^c(x) \le F_2^c(x)$ that $\bar{R}_1(t)/\lambda \ge \bar{R}_2(t)/\lambda$ for all $t \in [r, \tau]$. Then by (4.6) that

$$\bar{L}_1(\tau) - \bar{L}_1(r) \ge \bar{L}_2(\tau) - \bar{L}_2(r). \tag{A.3}$$

By the definition of r we have $\bar{Q}_1(r) = \bar{Q}_2(r)$ and $\bar{Z}_1(r) \le \bar{Z}_2(r) = 1$. It follows from (A.1) that $\bar{A}_1(r) - \bar{A}_2(r) < \delta$. Consequently, $r \ne \tau$. Then together with (A.2), (A.3) and the balance equation (4.7), we conclude

$$\begin{aligned} \bar{A}_1(\tau) - \bar{A}_2(\tau) &= \bar{A}_1(r) - \bar{A}_2(r) - [\bar{L}_1(\tau) - \bar{L}_1(r) - \bar{L}_2(\tau) \\ &\quad + \bar{L}_2(r)] - [\bar{Q}_1(\tau) - \bar{Q}_1(r) - \bar{Q}_2(\tau) \\ &\quad + \bar{Q}_2(r)] < \delta, \end{aligned}$$

which contradicts the definition of τ. So τ cannot be finite. Thus, we have proved that $\bar{A}_1(t) \le \bar{A}_2(t)$ for all $t \ge 0$. □

Definition 1. A blocked fluid model is specified by $\bar{Q}(t) \equiv 0$, equations (4.7)–(4.12), and

$$\bar{A}'(t) = \begin{cases} \lambda, & \bar{Z}(t) < 1, \\ \lambda \wedge \bar{S}'(t), & \bar{Z}(t) = 1. \end{cases}$$

Corollary 1. *Denote $\bar{A}_1(t)$ as the enter service process associated with a blocked fluid model. We have $\bar{A}_1(t) \le \bar{A}_2(t)$ for all $t \ge 0$, where $\bar{A}_2(t)$ is the corresponding process for an unblocked fluid model with the same arrival rate, service time distribution and initial state.*

Proof. The proof is almost identical to that of Lemma 1, we only point out the difference. We can use exactly the same argument leading to (A.2). Then, have $0 = \bar{Q}_1(t) \ge \bar{Q}_2(t)$ for all $t \in [r, \tau]$. It follows from (4.6) that $\bar{L}_2(\tau) - \bar{L}_2(r) = 0$. So the inequality (A.3) still holds. The rest of the argument is exactly same as that of Lemma 1. Thus we omit it. □

Limit of convolution. The following lemma is used in multiple places. The proof is quite standard, we omit for brevity.

Lemma 2. *If $f, g : [0, \infty) \to \mathbb{R}$ satisfy that $f(\infty) = \lim_{t\to\infty} f(t)$ and $\int_0^\infty |g(s)|ds < \infty$, then*

$$\lim_{t\to\infty} \int_0^t f(t-s)g(s)ds = f(\infty) \int_0^\infty g(s)ds.$$

Time shift of the fluid model. For any $\tau \geq 0$, denote $\left(\bar{\mathcal{Z}}_\tau(t), \bar{\mathcal{R}}_\tau(t)\right) = \left(\bar{\mathcal{Z}}(\tau+t), \bar{\mathcal{R}}(\tau+t)\right)$. The time shift for all the derived "status" quantities such as $\bar{Q}_\tau(\cdot)$, $\bar{R}_\tau(\cdot)$, $\bar{Z}_\tau(\cdot)$ and $\bar{X}_\tau(\cdot)$ are defined in the same way, e.g., $\bar{Q}_\tau(t) = \bar{Q}(\tau+t)$. However, let $\bar{A}_\tau(t) = \bar{A}(\tau+t) - \bar{A}(\tau)$ since $\bar{A}(t)$ records the "cumulative" amount of fluid that has entered service by time t.

Lemma 3. *Time-shifted fluid solution $\left(\bar{\mathcal{Z}}_\tau(t), \bar{\mathcal{R}}_\tau(t)\right)$ satisfies*

$$\begin{aligned}\bar{X}_\tau(t) &= \bar{\mathcal{Z}}(\tau)(C_t) + \bar{Q}(\tau)G^c(t) + \frac{\lambda}{\mu}\int_0^t H(\bar{Q}_\tau(t-s))dG_e(s) \\ &\quad + \int_0^t \bar{Q}_\tau(t-s)dG(s). \end{aligned} \tag{A.4}$$

Proof. Plugging (4.5) into (4.8) and applying integration-by-parts gives

$$\begin{aligned}\bar{\mathcal{Z}}(\tau)(C_t) &= \bar{\mathcal{Z}}(0)(C_{\tau+t}) + \frac{\lambda}{\mu}\int_t^{\tau+t} H(\bar{Q}(\tau+t-s))dG_e(s) \\ &\quad - \bar{Q}(\tau)G^c(t) + \bar{Q}(0)G^c(\tau+t) \\ &\quad + \int_t^{\tau+t} \bar{Q}(\tau+t-s)dG(s).\end{aligned}$$

So the right-hand side of (A.4) becomes

$$\begin{aligned}&\bar{\mathcal{Z}}(0)(C_{\tau+t}) + \bar{Q}(0)G^c(\tau+t) + \frac{\lambda}{\mu}\int_0^{\tau+t} H(\bar{Q}(\tau+t-s))dG_e(s) \\ &\quad + \int_0^{\tau+t} \bar{Q}(\tau+t-s)dG(s),\end{aligned}$$

which equals $\bar{X}(\tau+t)$ by (4.2). Thus (A.4) follows by applying the time shift definition. □

References

[1] R. Atar, C. Giat, N. Shimkin, The $c\mu/\theta$ rule for many server queues with abandonment, Oper. Res. 58 (5) (2010) 1427–1439.

[2] A. Bassamboo, R.S. Randhawa, Using estimated patience levels to optimally schedule customers. Technical Report, Northwestern University and USC, 2013.

[3] P. Billingsley, Convergence of Probability Measures, second ed., in: Wiley Series in Probability and Statistics, John Wiley & Sons Inc., New York, 1999.

[4] E. Hewitt, K. Stromberg, Real and Abstract Analysis, Springer-Verlag, New York, 1975.

[5] W. Kang, K. Ramanan, Fluid limits of many-server queues with reneging, Ann. Appl. Probab. 20 (6) (2010) 2204–2260.

[6] H. Kaspi, K. Ramanan, Law of large numbers limits for many-server queues, Ann. Appl. Probab. 21 (1) (2011) 33–114.

[7] A.A. Puhalskii, J.E. Reed, On many-server queues in heavy traffic, Ann. Appl. Probab. 20 (1) (2010) 129–195.

[8] J.E. Reed, The $G/GI/N$ queue in the Halfin–Whitt regime, Ann. Appl. Probab. 19 (6) (2009) 2211–2269.

[9] W. Whitt, Efficiency-driven heavy-traffic approximations for many-server queues with abandonments, Manage. Sci. 50 (10) (2004) 1449–1461.

[10] W. Whitt, Fluid models for multiserver queues with abandonments, Oper. Res. 54 (1) (2006) 37–54.

[11] J. Zhang, Fluid models of many-server queues with abandonment, Queueing Syst. 73 (2) (2013) 147–193.

[12] J. Zhang, J.G. Dai, B. Zwart, Law of large number limits of limited processor-sharing queues, Math. Oper. Res. 34 (4) (2009) 937–970.