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 Elsevier B.V. All rights reserved.

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 $\bar{\mathcal{R}}(t)(C_x)$ 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 $\bar{\mathcal{R}}(t)$. Introduce $\bar{R}(t) = \bar{\mathcal{R}}(t)(\mathbb{R})$, 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 $\bar{\mathcal{Z}}(t)(C_x)$ 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 $\bar{\mathcal{Z}}(t)$. 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 $(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t))$ evolves over time. For all $t \geq 0$,

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

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

Introduce $\bar{Z}(t) = \bar{\mathcal{Z}}(t)(C_0)$, the fluid content in service; and $\bar{Q}(t) = \bar{\mathcal{R}}(t)(C_0)$, 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 $(\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0))$ 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 $(\bar{\mathcal{R}}(0), \bar{\mathcal{Z}}(0))$. 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, $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is an *equilibrium state* of the fluid model (λ, F, G) if the solution to the fluid model with a valid initial state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ satisfies $(\bar{\mathcal{R}}(t), \bar{\mathcal{Z}}(t)) = (\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ for all $t \geq 0$. As characterized in Theorem 3.2 in [11], the state $(\bar{\mathcal{R}}_\infty, \bar{\mathcal{Z}}_\infty)$ is an equilibrium state of the fluid model (λ, F, G) if and only if it satisfies

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

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

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

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

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

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

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

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

then the convergence (3.5) holds.

4. Preliminary analysis

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

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