Available online at www.sciencedirect.com

ScienceDirect

Stochastic Processes and their Applications 124 (2014) 1436–1468

stochastic processes and their applications

www.elsevier.com/locate/spa

Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions

Alexander Walsh Zuñiga

Department of Industrial Engineering and Management, Technion Israel Institute of Technology, Haifa, Israel

Received 28 September 2012; received in revised form 25 November 2013; accepted 26 November 2013
Available online 4 December 2013

Abstract

This paper extends the works of Kang and Ramanan (2010) and Kaspi and Ramanan (2011), removing the hypothesis of absolute continuity of the service requirement and patience time distributions. We consider a many-server queueing system in which customers enter service in the order of arrival in a non-idling manner and where reneging is considerate. Similarly to Kang and Ramanan (2010), the dynamics of the system are represented in terms of a process that describes the total number of customers in the system as well as two measure-valued processes that record the age in service of each of the customers being served and the "potential" waiting times. When the number of servers goes to infinity, fluid limit is established for this triple of processes. The convergence is in the sense of probability and the limit is characterized by an integral equation.

MSC: primary 60K25; 60F17; 90B22; secondary 60G57; 60H20; 60J65

Keywords: Many-server queues; Fluid limits; Measure-valued processes; L^1-valued processes; Reneging; Abandonment

1. Introduction

We study a many server queueing system in which customers are served in a non-idling, First-Come-First-Served manner according to i.i.d. (independent, identically distributed) *service requirement*. Customers can leave the system, without getting service, once they have been waiting in the queue for more than their *patience time*, which are assumed to be also i.i.d. random

E-mail addresses: awalshz@gmail.com, awalshz@tx.technion.ac.il.

0304-4149/$ - see front matter
http://dx.doi.org/10.1016/j.spa.2013.11.008

variables. The objective is to obtain fluid approximations or functional strong laws of large numbers of characteristic functionals of the model, when N, the number of servers goes to infinity.

Many server systems were treated first in the seminal paper of Halfin and Whitt [5]. After that, many authors have succeeded relaxing the assumption imposed in [5], on one hand, considering general service requirement distributions and, on the other hand, incorporating the reneging option in their models. Existing work includes Kang and Ramanan [8], Kaspi and Ramanan [10], Mandelbaum and Momcilović [12], Puhalskii and Reed [15], Reed [16], Zhang [19] and many other references therein.

Generalizations of [5] are motivated by statistical analysis made by Brown et al. [3] which suggests that the service requirement distribution does not obey an exponential law. In [10], under the assumption that the distribution of the service requirement is absolutely continuous with respect to the Lebesgue measure, with density g satisfying a mild condition, fluid limits where obtained for a pair of processes $(\bar{X}^{(N)}, \bar{\nu}^{(N)})$. The first one is a nonnegative integer-valued process that represents the scaled total number of customers in system and the second one, a scaled measure-valued process that keeps track of the ages of customers in service.

With the application to call centers in mind, customer abandonment plays an important role and therefore must be considered. Garnett et al. [4] explain how the performance of certain systems are very sensitive to the impact produced by the impatience of customers. In this direction, the work in [10] has been extended in [8], including another measure-valued process $\bar{\eta}^{(N)}$ that keeps track of the "potential" waiting time of customers in the queue and they have obtained fluid limits adding this new process. The cumulative reneging process can be then expressed in terms of the triplet $(\bar{X}^{(N)}, \bar{\nu}^{(N)}, \bar{\eta}^{(N)})$ and the patience time must satisfy the same assumptions on the service requirement, that is, the existence of a density with respect to the Lebesgue measure.

Related to [8] and tracking the residual service and patience times instead of ages and potential waiting times, [19] obtained fluid limits approximations for the model studied in [8]. The approach of [19] avoids using martingales techniques, the fluid equations are different from those of [8], the functionals expressed in terms of the hazard rates are not taken into account. In this way, the fluid limits require weaker assumption on the service time distributions. The assumption for the service time distribution is continuity and for the patience time distribution is Lipschitz continuity.

Our work extends the results of [8,10,19] to the case of completely general service requirement distribution and continuous patience time distribution. We consider the same kind of reneging as in [8], that is, we assume that the queue is invisible to waiting customers, which is very suitable for call centers models. We obtain fluid limits for the triplet of process $(X^{(N)}, \nu^{(N)}, \eta^{(N)})$ of [8]. The fluid equations are very close to those considered in [8]. The difficulty arises in finding substitutes for the terms in the fluid equations of [8] depending on the densities of the distributions. For this purpose, we consider two sequences $(\bar{q}^{(N)})$ and $(\bar{p}^{(N)})$ of L^1-valued process (where L^1 is the set of integrable functions with respect to the Lebesgue measure on $\mathbb{R}_+$) which represent respectively for any time t and number of servers N, the densities of the measure $\int_0^t \bar{\nu}_s^{(N)} \mathrm{d}s$ and $\int_0^t \bar{\eta}_s^{(N)} \mathrm{d}s$ with respect to the Lebesgue measure on $\mathbb{R}_+$. These sequences always exist and the fluid equations are written in terms of their limits $\bar{q}$ and $\bar{p}$.

To take into account the processes $(\bar{q}^{(N)})$, $(\bar{p}^{(N)})$, $\bar{q}$ and $\bar{p}$ is very useful even in the case where the densities of service and patience time distributions exists. On one hand, it facilitates the analysis of convergence of the sequences $(\bar{\nu}^{(N)})$ and $(\bar{\eta}^{(N)})$ and provides convergence in probability of the process and not only weak convergence. On the other hand, it simplifies the proof of the uniqueness of the fluid equations. Indeed following [10], the delicate part is to