Available online at www.sciencedirect.com
SCIENCE DIRECT®
Operations Research Letters 33 (2005) 511–518
Operations Research Letters
www.elsevier.com/locate/orl

Heavy-traffic asymptotics for the single-server queue with random order of service

Bert Zwart[a, b, *]

[a] *Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*
[b] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*

Received 28 October 2003; accepted 15 September 2004
Available online 11 November 2004

Abstract

We consider the GI/GI/1 queue with customers served in random order, and derive the heavy-traffic limit of the waiting-time distribution. Our proof is probabilistic, requires no finite-variance assumptions, and makes the intuition provided by Kingman (Math. Oper. Res. 7 (1982) 262) rigorous.

MSC: 60K25

Keywords: Single-server queue; Joint queue length and workload distribution; Random order of service; Heavy traffic; Snapshot principle; State-space collapse

1. Introduction

In this paper we consider the GI/GI/1 queue where customers are served in random order. At the completion of a service, the server randomly takes one of the waiting customers into service. Classical papers on queues with random order of service (ROS) are those by Kingman [12], Palm [15] and Pollaczek [16]. Recently, the ROS discipline has received renewed interest. For example, collision resolution protocols in cable access networks operate in a manner quite similar to ROS; this was one motivation of the recent paper of Boxma et al. [4]. Other recent papers are by Flatto [7] and Borst et al. [2].

The present study is inspired by Boxma et al. [4]. That paper investigates several asymptotic properties of the GI/GI/1 ROS queue; in particular the tail of the steady-state waiting time W^{ROS} under heavy-tailed assumptions. They also consider the behavior of W^{ROS} when the system is in heavy traffic. Under the assumption of Poisson arrivals, it is shown in [4] that there exists a scaling function $\Delta(\rho)$ as $\rho \to 1$ such that

$$\Delta(\rho) W^{\mathrm{ROS}} \xrightarrow{\mathrm{d}} Y W^{\mathrm{FCFS}}. \tag{1}$$

* Corresponding author. Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.
E-mail address: zwart@win.tue.nl (B. Zwart).

0167-6377/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2004.09.006

Here, W^{FCFS} is the corresponding heavy-traffic limit of the workload (which is equal to the waiting time in FCFS) and Y is an exponential variable with mean 1, which is independent of W^{FCFS}. The derivation in [4] is based on Laplace-transform methods. A similar result was proved by Kingman [12] for the M/G/1 queue under more stringent conditions on the service-time distribution. Twenty years after the seminal paper [12], Kingman wrote an intriguing paper [14] in which he conjectured that an analog of (1) should hold in the GI/GI/1 queue.

The main goal of this paper is to settle that conjecture. More in particular, we present an insightful proof of (1), which does not need the assumption of Poisson arrivals. The proof is insightful, since it makes the heuristics outlined in [14] rigorous. As Kingman argues in his paper [14], if second moments of service times and inter-arrival times exist, the queue length (which is the same under ROS and FCFS) fluctuates on a time scale of $\mathrm{O}(1/(1-\rho)^2)$ when $\rho \to 1$. Since the waiting time is of the order $1/(1-\rho)$, the fluctuations of the queue length may be ignored. In the heavy-traffic literature, this is known as the *snapshot principle*. In Lemma 4.1 we make this precise and show that this line of thought is still valid if the finite-variance assumptions of Kingman [14] do not hold. Obtaining heavy-traffic limit theorems for queues with heavy tails is currently one of the main challenges in queueing theory; see the recent monograph of Whitt [18].

This paper is organized as follows. In Section 2 we introduce some notation, and state our heavy-traffic assumptions. Section 3 treats the heavy-traffic behavior of the joint queue length and workload distribution, which may be of independent interest, since these processes also concern FCFS. In particular, we show that the stationary queue length and waiting time in heavy traffic exhibit a form of *state-space collapse*—even in the heavy-tailed case. This complements recent process-level results in [18]. Our main result, an analog of Eq. (1), is stated and proved in Section 4.

2. Preliminaries

In this section we introduce some notation and state our assumptions. We consider a GI/GI/1 queue operating under the ROS discipline with load $\rho < 1$. W and Q are respectively the steady-state workload and queue length, as seen by an arriving customer. Thus, we consider the Palm-stationary workload and queue length. Note that the dynamics of the workload and queue length processes (and thus W and Q) are identical for the FCFS and ROS disciplines, as both service disciplines are work-conserving and non-preemptive. In particular, W can be identified with the steady-state waiting time under FCFS.

Since we are interested in the performance of the GI/GI/1 ROS queue in heavy traffic, we will let $\rho \to 1$. For this purpose, it is convenient to index all random variables by r; in the rth system, the inter-arrival times and service times are respectively given by the mutually independent i.i.d. sequences $A_{i,r}, i \geqslant 1$ and $B_{i,r}, i \geqslant 1$. Define further $S^A_{n,r} = A_{1,r} + \cdots + A_{n,r}$, $S^B_{n,r} = B_{1,r} + \cdots + B_{n,r}$, $N^A_r(t) = \max\{n : S^A_{n,r} \leqslant t\}$, $N^B_r(t) = \max\{n : S^B_{n,r} \leqslant t\}$, and $Y_r(t) = S^B_{N^A_r(t),r}$. Note that $Y_r(t)$ is the amount of work arrived into the system between time 0 and t.

Our first assumption states our heavy-traffic condition, and ensures that all processes defined above satisfy triangular array versions of the functional weak law of large numbers. A sufficiently general formulation of such a result is given in Lemma A.2 in Gromoll et al. [10]; see also Section 1.5(c) of Durrett [6].

Before we state this assumption, we introduce some further notation which is used throughout this paper. If a limit is taken, it is always the limit as $r \to \infty$, unless stated otherwise. A similar statement applies to order symbols. With $\xrightarrow{\mathrm{d}}$ we mean convergence in distribution.

Assumption 1 (*Heavy traffic*). There exist two mutually independent i.i.d. sequences $A_i, i \geqslant 1$ and $B_i, i \geqslant 1$ with $E[A_1] = E[B_1]$ such that as $r \to \infty$, $A_{i,r} \xrightarrow{\mathrm{d}} A_i$, $B_{i,r} \xrightarrow{\mathrm{d}} B_i$, $E[A_{1,r}] \to E[A_1]$, $E[B_{1,r}] \to E[B_1]$, and $r(1-\rho_r) = r(1 - E[B_r]/E[A_r]) \to 1$. Furthermore, we assume that $E[A_{i,r} I(A_{i,r} > r)] \to 0$, $E[B_{i,r} I(B_{i,r} > r)] \to 0$.

Our second assumption is concerned with the existence of a heavy-traffic limit for the stationary (w.r.t. customer arrivals) workload W_r. This is a topic which has been well studied. Three approaches to the