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

TU/e technische universiteit eindhoven

SPOR-Report 2003-26

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

A.P. Zwart

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven, October 2003
The Netherlands

/department of mathematics and computing science

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, Statistics and Operations research
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariat: Main Building 9.10
Telephone: + 31 40 247 3130
E-mail: wscosor@win.tue.nl
Internet: http://www.win.tue.nl/math/bs/cosor.html

ISSN 1567-5211

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

Bert Zwart

Department of Mathematics & Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
zwart@win.tue.nl

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

October 28, 2003

Abstract

We consider the waiting time distribution of the $GI/GI/1$ queue where customers are served in random order; inter-arrival and service times may have finite or infinite variance. Our main result shows that the waiting time in heavy traffic can be written as a product of two random variables. Our proof is based on the intuitively appealing fact that in heavy traffic, the queue length stays constant during the sojourn time of a customer. For the special finite variance case, our result settles a conjecture of Kingman (1982).

2000 Mathematics Subject Classification: 60K25.
Keywords & Phrases: 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 Kingman [10], Palm [13] and Pollaczek [14]. 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 [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

$$\Delta(\rho) W^{ROS} \xrightarrow{d} Y W^{FCFS}. \tag{1.1}$$

Here, Y is an exponential variable with mean 1, and W^{FCFS} is the corresponding heavy traffic limit of the workload (which is equal to the waiting time in FCFS). The derivation in [4] is based on Laplace-transform methods. A similar result was proven by Kingman [10] for the $M/G/1$ queue under more stringent conditions on the service-time distribution. Twenty years after the seminal paper [10], Kingman wrote an intriguing paper [12] in which he conjectured that an analogue of (1.1) 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 (1.1), which does not need the assumption of Poisson arrivals. The proof is insightful, since it makes the heuristics outlined in Kingman [12] rigorous. As Kingman argues in his paper [12], 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 $O(1/(1-\rho)^2)$ 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 [12] 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 [16].
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 Whitt [16]. Our main result, an analogue of Equation (1.1), is stated and proven in Section 4.

2 Preliminaries: The workload in heavy traffic

In this section we consider the steady-state (w.r.t. customer arrivals) workload W and queue length Q - as seen by an arriving customer. Observe that the workload and queue length processes are identical for the FCFS and ROS disciplines. Thus, we can represent the steady-state workload as follows:

$$W = \sup_{n \geq 0} S_n,$$

here $S_n = \sum_{i=1}^{n} X_i, n \geq 1$, and $X_i = B_i - A_i$, where B_i and $A_i, i \geq 1$, are i.i.d. sequences of service and inter-arrival times. Write $\rho = E[B_1]/E[A_1]$. 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 r-th system, inter-arrival times and service times are given by $A_{i,r}$ and $B_{i,r}$. Define the process

$$S_r(t) = S_{r,\lfloor t \rfloor}.$$

Our first assumption is that the input process $S_r(t)$ satisfies a functional central limit theorem, and that $\rho_r \to 1$. More precisely, we assume

Assumption 2.1 *(Heavy traffic.) There exist normalizing constants c_r and $d_r = rc_r$ such that the normalized process*

$$\hat{S}_r(t) = c_r^{-1}(S_r(d_r t) - \lfloor d_r t \rfloor (E[B_{1,r}] - E[A_{1,r}]))$$

converges in $D[0,\infty)$ *to an* $\alpha-$*stable Lévy process* $X(t), t \geq 0$, $\alpha \in (1,2]$. *Moreover,* $A_{i,r} \xrightarrow{d} A_i$, $B_{i,r} \xrightarrow{d} B_i$, $E[A_{1,r}] \to E[A_1]$, $E[B_{1,r}] \to E[B_1] = E[A_1]$ *in such a way that* $r(1-\rho_r) = r(1 - E[B_r]/E[A_r]) \to 1$.

Throughout this paper (in particular in Assumption 2.1 above) we use the following notational conventions. If a limit is taken (denoted by $\to$), it is always the limit as $r \to \infty$, unless stated otherwise. A similar statement applies to order symbols. With $\xrightarrow{d}$ we mean convergence in distribution. In Assumption 2.1 given above, convergence in $D[0,\infty)$ is w.r.t. the (standard) Skorokhod J_1-topology. More information about this space and its various topologies can be found in Billingsley [5] and Whitt [16].
Two main examples in which the above condition is satisfied are:

Example L: The light-tailed (finite variance) case
Let $B_{i,r} = B_i$, and $A_{i,r} = A_i/\rho_r$. Suppose that $E[A_1] = E[B_1]$, and $E[A_1^2], E[B_1^2] < \infty$. In this case, Assumption 2.1 is satisfied, and one can choose $c_r = r, d_r = r^2$. The limiting process is Brownian motion (i.e. $\alpha = 2$).

Example H: Heavy-tailed service times
Let again $B_{i,r} = B_i$, $A_{i,r} = A_i/\rho_r$, and $E[A_1] = E[B_1]$. Moreover, let $P(B_1 > x) = L(x)x^{-\nu}, 1 < \nu < 2$, with L slowly varying, and assume that $E[A_1^\eta] < \infty$ for some $\eta > \nu$. In this case, Assumption 2.1 is satisfied with $\alpha = \nu$, and scaling constants c_r chosen such that $c_r P(B_1 > c_r) \sim 1/r$ and $d_r = c_r r$. Thus, c_r is regularly varying of index $1/(\nu-1)$ and d_r is regularly varying with index $\nu/(\nu-1)$.

Other sufficient conditions for Assumption 2.1 to hold are provided in, for example, Resnick & Samorodnitsky [15], and Whitt [16].
Consider now the workload W_r. Since

$$W_r = \sup_{t\geq 0} S_r(t),$$

we have

$$\begin{aligned} \hat{W}_r &:= c_r^{-1} W_r \\ &= \sup_{t\geq 0} \hat{S}_r(t) + c_r^{-1}[d_r t](E[B_{1,r}] - E[A_{1,r}]). \end{aligned}$$

Since, in view of Assumption 2.1, $\hat{S}_r(t) \xrightarrow{d} X(t)$, and $c_r^{-1}[d_r t](E[B_{1,r}] - E[A_{1,r}]) \to -t$, one is tempted to conclude that $\hat{W}_r \to \sup_{t>0}(X(t) - t)$ as $r \to \infty$. However, this is not a trivial matter. A reason for this is that the functional $\sup_{t>0}$ is not continuous in $D[0,\infty)$. Thus, one can not apply the continuous mapping theorem. In general, additional regularity conditions are needed. Since the focus of this work is on the ROS policy, we will just *assume* that the problem of establishing a heavy-traffic limit theorem for the workload is settled.

Assumption 2.2 *There exists a random variable* W^* *with a continuous distribution such that*

$$\hat{W}_r \xrightarrow{d} W^* \stackrel{d}{=} \sup_{t\geq 0} X(t) - t.$$

This assumption turns out to be quite natural in the next two sections, and is valid in our two motivating examples given above:

Example L (continued)
In the finite variance setting as given above, Kingman [9, 11] has shown that

$$P((1/r)W_r > x) \to P(W^* > x) = \mathrm{e}^{-cx}, \tag{2.2}$$

with $c = 2E[A_1]/(Var[A_1] + Var[B_1])$; see Chapter X.7 of Asmussen [1] for a textbook treatment.

Example H (continued)
This example falls in the framework of Resnick & Samorodnitsky [15] (their conditions (A)–(C) are easily shown to hold for our example). In particular, using Corollary 2.2 of [15] we obtain

$$P(c_r^{-1}W_r > x) \to P(W^* > x) = \sum_{n=0}^{\infty} \frac{(-a)^n}{\Gamma(1+n(\nu-1))} x^{n(\nu-1)}, \tag{2.3}$$

with $a = (\nu-1)/\Gamma(2-\nu)$, and $\Gamma(\cdot)$ the Gamma function. The sum on the right-hand side is known as the *Mittag Leffler function*; see Furrer *et. al* [8] for related result in a risk model. Note that the Laplace-Stieltjes transform of W^* is given by

$$E[\mathrm{e}^{-sW^*}] = \frac{a}{a+s^{\nu-1}}. \tag{2.4}$$

Under more stringent assumptions, the result (2.3) is also obtained by Boxma & Cohen [3].

3 The joint workload and queue-length distribution

The present section focuses on the steady-state (again w.r.t. arrival epochs) queue length Q_r in heavy traffic. More precisely, we consider the *joint* distribution of Q_r and W_r as $r \to \infty$.
Let $\hat{Q}_r = c_r^{-1}Q_r$ be the rescaled queue length. Furthermore, let B_r^e be the residual service time of the customer in service. (Put $B_r^e = 0$ if $Q_r = 0$). Define $\hat{B}_r^e = c_r^{-1}B_r^e$. We have the following identity:

$$W_r = B_r^e + \sum_{i=1}^{Q_r} B_{i,r}. \tag{3.5}$$

In this expression, the $B_{i,r}$ are independent of the pair (Q_r, B_r^e). Since it is intuitively clear that $\hat{B}_r^e \to 0$, one is tempted to conclude from Equation (3.5) that $\hat{Q}_r$ and $\hat{W}^r$ are equal up to a multiplicative constant when $r \to \infty$.
However, it is not easy to make this rigorous. For example, a problem is that Q_r and B_r^e are not independent. Therefore, we need another representation for the joint distribution of W_r and Q_r. This representation is given by the following lemma, which may be of independent interest.

Lemma 3.1 *For any $k \geq 1, y \geq 0$,*

$$P(Q_r \geq k; W_r > y) = P(W_r + B_r > T_{r,k}; \max(W_r + S_k, \max_{1 \leq i \leq k} S_i) > y).$$

Proof
This result, as well as its proof, is an extension of Theorem X.4.3 of Asmussen (2003). Let $Q_{r,n}$ be the queue length seen by the $n-$th arrival in the $r-$th system and similarly, let $W_{r,n}$ be the amount of work in the system right before the $n-$th arrival.
As in Asmussen (2003), we observe that

$$\{Q_{r,n+k} \geq k\} = \{T_{r,n+k} \leq T_{r,n} + W_{r,n} + B_{r,n}\},$$

where $T_{r,n} = A_{r,1} + \ldots + A_{r,n}$. We now express $W_{r,n+k}$ in terms of $W_{r,n}$. From Lindley's recursion, it readily follows that

$$\{W_{r,n+k} > y\} = \{\max(W_{r,n} + S_{r,n+k} - S_{r,n}, \max_{1 \leq i \leq k} S_{r,n+i} - S_{r,n}) > y\}.$$

The result now follows by combining the two events, taking probabilities, letting $n \to \infty$, and observing that the vector $(Q_{r,n}, W_{r,n})$ weakly converges to (Q_r, W_r).

□

We are now ready to present the main result of this section.

Proposition 3.1 *Suppose Assumptions 2.1 and 2.2 hold. Then*

$$(\hat{W}_r, \hat{Q}_r, \hat{B}_r^e) \to (1, 1/E[A_1], 0)W^*.$$

Proof
First, we prove joint convergence of $(\hat{Q}_r, \hat{W}_r)$. From Lemma 3.1, we observe that

$$P(\hat{Q}_r > x; \hat{W}_r > y) = P(W_r + B_r > T_{r,[xc_r]}; \max(W_r + S_{r,[xc_r]}, \max_{1 \leq i \leq [xc_r]} S_{r,i}) > y).$$

Because of the strong law of large numbers we have (i) $T_{r,[xc_r]}/c_r \to xE[A]$ a.s.; (ii) $S_{r[xc_r]} \to 0$ a.s.; (iii) $\max_{1 \leq i \leq [xc_r]} S_{r,i} \to 0$ a.s. . Combining (i), (ii) and (iii) we obtain

$$P(\hat{Q}_r > x; \hat{W}_r > y) \to P(W^* > E[A]x; W^* > y).$$

This implies convergence in distribution of $(\hat{Q}_r, \hat{W}_r)$ to $(1/E[A], 1)W^*$.
It remains to consider the convergence of $\hat{B}_r^e$. For this, we use (3.5) to obtain

$$\hat{B}_r^e = \hat{W}_r - c_r^{-1} \sum_{i=1}^{c_r \hat{Q}_r} B_{i,r}.$$

Using the established convergence for $(\hat{W}_r, \hat{Q}_r)$ and the strong law of large numbers, it is straightforward to show that

$$(\hat{W}_r, c_r^{-1} \sum_{i=1}^{c_r \hat{Q}_r} B_{i,r}) \xrightarrow{d} (1, 1)W^*.$$

The desired statement is now immediate.

□

4 Main result

Let W_r^{ROS} be the steady-waiting time of a customer in the $GI/GI/1$ ROS queue. Invoking Assumption 2.1, we can define the scaled waiting time

$$\hat{W}_r^{ROS} = c_r^{-1} W_r^{ROS}.$$

In this section we prove the following result. Let Y be an exponential random variable with mean 1, which is independent of everything else.

Theorem 4.1 *Assume that Assumptions 2.1 and 2.2 hold. Then*

$$\hat{W}_r^{ROS} \xrightarrow{d} W_*^{ROS} = YW^*.$$

Before we give a proof of Theorem 4.1, we give two applications:
Example L (continued)
Recall that $B_{i,r} = B_i$, $A_{i,r} = A_i/\rho_r$, $E[A_1] = E[B_1]$, and $E[A_1^2], E[B_1^2] < \infty$. Combining Theorem 4.1 and (2.2) we obtain

$$P((1/r)W_r^{ROS} > x) \to P(YW^* > x) = \int_0^\infty \mathrm{e}^{-u-cx/u}\mathrm{d}u = 2\sqrt{cx}K_1(2\sqrt{cx}), \qquad (4.6)$$

with $K_1(x)$ the modified Bessel function of the second kind.

Example H (continued)
Recall that the setting in this example is $B_{i,r} = B_i$, $A_{i,r} = A_i/\rho_r$, and $E[A_1] = E[B_1]$. $P(B_1 > x) = L(x)x^{-\nu}, 1 < \nu < 2$, and $E[A_1^\eta] < \infty$ for some $\eta > \nu$. Combining Theorem 4.1 with (2.3) we obtain

$$P(c_r^{-1}W_r^{ROS} > x) \to \int_0^\infty \mathrm{e}^{-x/y}\mathrm{d}P(W^* \le y) = \int_0^\infty \mathrm{e}^{-y}P(W^* > x/y)\mathrm{d}y,$$

where the distribution of W^* is given by the right-hand side of (2.3). If ν is irrational, one can rewrite the above integral (using $\Gamma(1-y) = \pi/(\Gamma(y)\sin y)$) to obtain

$$P(c_r^{-1}W_r^{ROS} > x) \to \sum_{n=0}^\infty \frac{\pi(-a)^n}{\sin(\pi n(\nu-1))\Gamma(n(\nu-1))\Gamma(1+n(\nu-1))} x^{n(\nu-1)}.$$

It is also possible to obtain heavy traffic approximations of W_r in the heavy-tailed case when the condition $E[A^\eta] < \infty$ is violated, see e.g. Cohen [6]. These results can be combined with Theorem 3.1 to get the corresponding heavy-traffic limit for W_r^{ROS}.

We now turn to a proof of Theorem 4.1. Our proof relies on the following crucial lemma, for which we need to introduce some more notation. Let $Q_r(t)$ be the number of customers at time t in the $r-$th system, and take $Q_r(0) = Q_r$. Define also the scaled queue length process $\hat{Q}_r(t) = c_r^{-1}Q_r(t)$.

Lemma 4.1 *Assume that Assumptions 2.1 and 2.2 hold. Then, for every fixed $M \in (0,\infty)$ and every $\gamma > 0$,*

$$P(\sup_{t\in[0,Mc^r]} |\hat{Q}_r(t) - \hat{Q}_r(0)| > \gamma) \to 0.$$

In words, this lemma states that the scaled queue length process does not fluctuate much between time 0 and Mc_r. Since waiting times will be shown to be of order c_r, this lemma shows that the snapshot principle as desribed in the introduction is indeed valid.

Proof

Fix M and take an arbitrary $\eta > 0$. Note that

$$\begin{aligned} & P(\sup_{t\in[0,Mc^r]} |\hat{Q}_r(t) - \hat{Q}_r(0)| > \gamma) \\ \leq\ & P(\hat{Q}_r(t) \leq \eta) + P(\sup_{t\in[0,Mc^r]} |\hat{Q}_r(t) - \hat{Q}_r(0)| > \gamma, \hat{Q}_r(t) > \eta). \end{aligned}$$

Note that the first probability converges to 0 for any $\eta > 0$, using Proposition 2.1 and the fact that $P(Q^* > 0) = P(W^* > 0) = 1$.

We now proceed by using a variation of the argument made in Lemma 3.1 of [4].

Consider the event $E_{r,\delta}$ given by

$$E_{r,\delta} = \{Q_r(t) \in [(1-\delta)Q_r(0) - (1-\rho_r+\delta)t, (1+\delta)Q_r(0) - (1-\rho_r-\delta)t], 0 < t < Mc_r\}.$$

By the strong law of large numbers, for every $\delta > 0$ sufficiently small (w.r.t. η) there exists a $r^* = r(\delta, \eta)$ such that

$$P(E_{r,\delta} \mid \hat{Q}_r(t) > \eta) > 1 - \delta, \qquad r \geq r^*.$$

It is not difficult to show that, under $E_{r,\delta}$,

$$\sup_{t\in[0,Mc^r]} |\hat{Q}_r(t) - \hat{Q}_r(0)| \leq \delta(\hat{Q}_r(0) + M).$$

Thus, we conclude that for every $\eta > 0$ and for every $\delta << \eta$, $\delta > 0$,

$$\begin{aligned} & P(\sup_{t\in[0,Mc^r]} |\hat{Q}_r(t) - \hat{Q}_r(0)| > \gamma) \\ \leq\ & P(\hat{Q}_r(t) \leq \eta) + P(\sup_{t\in[0,Mc^r]} |\hat{Q}_r(t) - \hat{Q}_r(0)| > \gamma, \hat{Q}_r(t) > \eta) \\ \leq\ & P(\hat{Q}_r(0) \leq \eta) + P(E_\delta^c) + P(\delta(\hat{Q}_r(0) + M) > \gamma) \\ \to\ & P(Q^* \leq \eta) + 0 + P(Q^* + M > \gamma/\delta), \end{aligned}$$

as $r \to \infty$. Finally, let first $\delta \downarrow 0$ and then $\eta \downarrow 0$ to complete the proof.

□

We are now ready to prove Theorem 4.1.

The idea behind the proof is simple: Lemma 4.1 implies that, on the time scale c_r, the queue length process hardly changes in heavy traffic (i.e., as $r \to \infty$). Thus, a snapshot principle holds: in heavy traffic, a customer does not see any change in the queue lenght during its waiting time. In the ROS context, this means that the probability of being the next customer in service is $1/Q_r(0)$ throughout its waiting time. This implies that its waiting time is approximately given by

$$W_r^{ROS} \approx \sum_{i=1}^{G(1/Q_r(0))} B_{r,i},$$

with $G(p)$ a geometrically distributed random variable with rate p. Since $pG(p)$ weakly converges to the exponentially distributed random variable Y, this implies that, as $r \to \infty$,

$$\hat{W}_r^{ROS} \approx YE[B]W^*/E[A] = YW^*.$$

Proof of Theorem 4.1
According to Lemma 4.1 we have

$$P(\hat{W}_r^{ROS} > y) = P(\hat{W}_r^{ROS} > y; A_r(\gamma, y)) + o(1) \tag{4.7}$$

as $r \to \infty$, where the event $A_r(\gamma, y)$ is given by

$$A_r(\gamma, y) = \{|\hat{Q}_r(t) - \hat{Q}_r(0)| \leq \gamma, 0 \leq t \leq yc_r\}.$$

Under this event, it is possible to get tractable lower and upper bounds for W_r^{ROS}. The remainder of this proof consists of deriving these bounds and showing that they behave similarly when $r \to \infty$.
Let $G(p), p \in (0,1)$ be a family of geometrically distributed random variables with success parameter p, independent of everything else. Define the random variable

$$p_{\gamma,r} = \frac{1}{\max\{(Q_r(0) - \gamma c_r), 1\}}.$$

Let $G_{\gamma,r}$ be a "mixed" geometric random variable, i.e.,

$$P(G_{\gamma,r} > u) = \int_0^1 P(G(p) > u) \mathrm{d}P(p_{\gamma,r} \leq p).$$

Then, because of the nature of the ROS discipline, the following inequalities are valid:

$$P(\hat{W}_r^{ROS} > y; A_r(\gamma, y)) \leq P(\hat{B}_r^e + c_r^{-1} \sum_{i=1}^{G_{\gamma,r}} B_{i,r} > y; A_r(\gamma, y)), \tag{4.8}$$

$$P(\hat{W}_r^{ROS} > y; A_r(\gamma, y)) \geq P(\hat{B}_r^e + c_r^{-1} \sum_{i=1}^{G_{-\gamma,r}} B_{i,r} > y; A_r(-\gamma, y)). \tag{4.9}$$

Since $G_{\gamma,r} \to \infty$ as $r \to \infty$, we can simplify the above lower and upper bounds using the strong law of large numbers for $\sum_{i=1}^n B_{i,r}$. Combining this once more with Lemma 4.1 we then obtain, for each $\epsilon > 0$, the following upper bound from (4.8):

$$P(\hat{W}_r^{ROS} > y; A_r(\gamma, y)) \leq P(\hat{B}_r^e + c_r^{-1} G_{\gamma,r} E[B](1+\epsilon) > y) + o(1). \tag{4.10}$$

The lower bound (4.9) can be simplified even further since $\hat{B}_r^e \geq 0$:

$$P(\hat{W}_r^{ROS} > y; A_r(\gamma, y)) \geq P(c_r^{-1} G_{-\gamma,r} E[B](1-\epsilon) > y) + o(1). \tag{4.11}$$

It thus suffices to consider the weak convergence properties of the random variable $\hat{G}_{\gamma,r} = c_r^{-1} G_{\gamma,r}$ for any fixed γ in a neighborhood of 0. More precisely, we need to investigate the joint convergence properties of the vector $(\hat{B}_r^e, \hat{G}_{\gamma,r})$.
This is done in the following Lemma, which is proven after having finished the proof of Theorem 4.1. Define $Q^* = W^*/E[A] = W^*/E[B]$.

Lemma 4.2 *If Assumptions 2.1 and 2.2 are satisfied, then*

$$(\hat{B}_r^e, \hat{G}_{\gamma,r}) \xrightarrow{d} (0, Y(Q^* + \gamma)^+).$$

Combining Lemma 4.2 with (4.7) and (4.10) we obtain, for every $\gamma > 0$,

$$\limsup_{r\to\infty} P(\hat{W}_r^{ROS} > y) \leq P(Y(Q^* + \gamma)E[B](1 + \epsilon) > y).$$

Since this is true for any choice of $\gamma, \epsilon > 0$ we obtain,

$$\limsup_{r\to\infty} P(\hat{W}_r^{ROS} > y) \leq P(YW^* > y). \tag{4.12}$$

Similarly, from Lemma 4.2 with (4.7) and (4.10) we obtain, for every $\gamma > 0$,

$$\liminf_{r\to\infty} P(\hat{W}_r^{ROS} > y) \geq P(Y(Q^* - \gamma)^+ E[B](1 - \epsilon) > y).$$

Since this is true for any choice of $\gamma, \epsilon > 0$, we obtain

$$\liminf_{r\to\infty} P(\hat{W}_r^{ROS} > y) \geq P(YW^* > y). \tag{4.13}$$

Combining (4.12) and (4.13) completes the proof.

□

It remains to prove Lemma 4.2.

Proof of Lemma 4.2

Take $\delta > 0$, and consider the probability

$$P(\hat{B}_r^e < \delta, c_r^{-1} G(p_{\gamma,r}) > x) = \int_{y=0}^{c_r} (1 - y/c_r)^{\lfloor x c_r \rfloor} \mathrm{d}P(c_r p_{\gamma,r} \leq y; \hat{B}_r^e < \delta).$$

Since $(\hat{B}_r^e, p_{\gamma,r}) \xrightarrow{d} (0, 1/(Q^* + \gamma)^+)$, we immediately obtain, by Fatou's lemma,

$$\liminf_{r\to\infty} P(\hat{B}_r^e < \delta, c_r^{-1} G(p_{\gamma,r}) > x) \geq P(Y(Q^* + \gamma)^+ > x).$$

To get an upper bound, note that $(1 - a)^b \leq \mathrm{e}^{-ab}$ if $a \in (0, 1]$. In addition, take $\epsilon > 0$, and note that $c_r^{-1} \leq \epsilon$ for r large enough. Thus, for r large enough, we conclude that

$$P(\hat{B}_r^e < \delta, c_r^{-1} G(p_{\gamma,r}) > x) \leq \int_{y=0}^{c_r} \mathrm{e}^{-(x-\epsilon)y} \mathrm{d}P(c_r p_{\gamma,r} \leq y; \hat{B}_r^e < \delta).$$

From the weak convergence of $(\hat{B}_e^r, c_r p_{\gamma,r})$ we finally conclude that, for every $\epsilon > 0$,

$$\limsup_{r\to\infty} P(\hat{B}_r^e < \delta, c_r^{-1} G(p_{\gamma,r}) > x) \leq P(Y(Q^* + \gamma)^+ > x - \epsilon).$$

Letting $\epsilon \downarrow 0$ completes the proof.

□

References

[1] Asmussen, S. (2003). *Applied Probability and Queues.* 2nd edition. Springer, New York.

[2] Borst, S.C., Boxma, O.J., Núñez-Queija, R., Morrison, J. (2003). The equivalence between processor sharing and service in random order. *Operations Research Letters* **31**, 254–262.

[3] Boxma, O.J., Cohen, J.W. (1999). Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. *Queueing Systems* **33**, 177–204.

[4] Boxma, O.J., Foss, S., Lasgouttes, J.M., Núñez-Queija, R. (2002). Waiting time asymptotics in the single-server queue with service in random order. *Queueing Systems*, to appear.

[5] Billingsley, P. (1968). *Convergence of Probability measures.* wiley, New York.

[6] Cohen, J.W. (1997). Heavy-traffic limit theorems for the heavy-tailed $GI/G/1$ queue. Report PNA-R9719, CWI, Amsterdam.

[7] Flatto, L. (1997). The waiting time distribution for the random order service $M/M/1$ queue. *Annals of Applied Probability* **7**, 382–409.

[8] Furrer, H., Michna, Z., Weron, A. (1997). Stable Lévy motion approximation in collective risk theory. *Insurance: Mathematics and Economics* **20**, 97–114.

[9] Kingman, J.F.C. (1961). The single server queue in heavy traffic. *Proceedings of the Cambridge Mathematical Society* **57**, 902–904.

[10] Kingman, J.F.C. (1962). On queues in which customers are served in random order. *Proceedings of the Cambridge Philosophical Society* **58**, 79–91.

[11] Kingman, J.F.C. (1965). The heavy traffic approximation in the theory of queues. In: Smith, W.L., Wilkinson, W.E. (editors): *Proceedings of the Symposium on Congestion Theory,* 137–169. Univ. North Carolina Press, Chapel Hill, N.C.

[12] Kingman, J.F.C. (1982). Queueing disciplines in heavy traffic. *Mathematics of Operations Research* **7**, 262–271.

[13] Palm, C. (1957). Waiting times with random served queue. *Tele1* 1–107. (English edition; original from 1938).

[14] Pollaczek, F. (1946). La loi d'attente des appels téléphoniques. *C.R. Acad. Sci. Paris* **222**, 353–355.

[15] Resnick, S., Samorodnitsky, G. (2000). A heavy traffic limit theorem for workload processes with heavy tailed service requirements. *Management Science* **46**, 1236–1248.

[16] Whitt, W. (2002). *Stochastic-Process Limits.* Springer, New York.