Waiting time asymptotics in the single server queue with service in random order

Onno Boxma, Serguei Foss, Jean-Marc Lasgouttes, Rudesindo Núñez Queija

Waiting time asymptotics in the single server queue with service in random order

O.J. Boxma [*†] S.G. Foss [‡] J.-M. Lasgouttes [§] R. Núñez Queija[*†]

July 18, 2012

Abstract

We consider the single server queue with service in random order. For a large class of heavy-tailed service time distributions, we determine the asymptotic behavior of the waiting time distribution. For the special case of Poisson arrivals and regularly varying service time distribution with index $-\nu$, it is shown that the waiting time distribution is also regularly varying, with index $1-\nu$, and the pre-factor is determined explicitly.

Another contribution of the paper is the heavy-traffic analysis of the waiting time distribution in the $M/G/1$ case. We consider not only the case of finite service time variance, but also the case of regularly varying service time distribution with infinite variance.

Keywords: single server queue, service in random order, heavy-tailed distribution, waiting time asymptotics, heavy-traffic limit theorem.

Acknowledgement: J.-M. Lasgouttes did most of his research for the present study while spending a sabbatical at EURANDOM in Eindhoven. O.J. Boxma and S.G. Foss gratefully acknowledge the support of INTAS, project 265 on "The mathematics of stochastic networks".

1 Introduction

We consider a single server queue that operates under the Random Order of Service discipline (ROS; also SIRO = Service In Random Order): At the completion of a service, the server randomly takes one of the waiting customers into service. Research on the ROS discipline has a rich history, inspired by its natural occurrence in several problems in telecommunications. The $M/M/1$ queue with ROS was studied by Palm [25], Vaulot [28] and Pollaczek [26, 27]. Burke [11] derived the waiting time distribution in the $M/D/1$ case. An expression for the (Laplace-Stieltjes transform of the) waiting time distribution for the $M/G/1$ case was obtained by Kingman [20] and Le Gall [22]; the former author also studied the *heavy-traffic* behavior of the waiting time distribution, when the service times have a finite variance. Quite

[*]Department of Mathematics & Computer Science and EURANDOM; Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

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

[‡]Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Riccarton, Edinburgh, EH14 4AS UK

[§]INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex, France

recently, Flatto [15] derived detailed tail asymptotics of the waiting time in the $M/M/1$ case. As pointed out in Borst et al. [6], this immediately yields detailed tail asymptotics of the sojourn time in the $M/M/1$ queue with Processor Sharing, because these two quantities are closely related in a single server queue with exponential service times.

In the present study, we are also interested in waiting time tail asymptotics of single server queues with ROS. However, here we concentrate on the case of *heavy-tailed* service time distributions. The motivation for this study is twofold. Firstly, an abundance of measurement studies regarding traffic in communication networks like local area networks and the Internet has made it clear that such traffic often has heavy-tailed characteristics. It is therefore important to investigate the impact of such traffic on network performance and to determine whether possibly adverse effects can be overcome by employing particular traffic management schemes. One possibility is to modify the 'service discipline' (i.e., scheduling mechanism); this may lead to a significant change in performance [7].

Secondly, in real life there are many situations in which service is effectively given in random order. Our own interest in ROS was recently revived in a joint project with Philips Research concerning the performance analysis of cable access networks. Collision resolution of user requests for access to the common transmission channel is being handled by a Capetanakis-Tsybakov-Mikhailov type tree protocol [3]. That collision resolution protocol handles the requests in an order that is quite close to ROS [9].

We now present an outline of the organization and main results of the paper. Section 2 contains preliminary results on the busy period and waiting time tail behavior in the $GI/G/1$ queue with a non-preemptive and non-idling service discipline. They are used in Section 3 to study the waiting time tail for the $GI/G/1$ queue with service in random order. The tail of the service time distribution is assumed to be in the class $\mathcal{L} \bigcap \mathcal{D}$. This class contains the class of regularly varying distributions; these two classes, and others, are briefly discussed in Appendix A. We sketch a probabilistic derivation of the asymptotic behavior of the waiting time distribution, deferring a detailed derivation to Appendix C. For large x, $\mathsf{P}(W_{\rm ROS} > x)$ is written as a sum of four terms, each of which has a probabilistic interpretation. These interpretations are based on the knowledge that, for sums of independent random variables with a subexponential distribution, the most likely way for the sum to be very large is that *one of the summands* is very large (similar ideas were developed in [2] for a class of stochastic networks – see the so-called 'Typical Event Theorem' there). For example, the first of the four terms equals ρ times the probability that a residual service time is larger than x, ρ denoting the traffic load. The probabilistic interpretation is that one possibility for the waiting time of a tagged customer to be larger than some large value x is, that the residual service time of the customer in service upon his arrival exceeds x. The other three terms are more complicated, taking into account possibilities like: A customer with a very large service time has already left when the tagged customer arrived, but it has left a very large number of customers behind — and the tagged customer has to wait for many of those (and newly arriving) customers.

In the subsequent sections we restrict ourselves to the case of Poisson arrivals. In the case of an $M/G/1$ queue with regularly varying service time distribution, we are able to obtain detailed tail asymptotics for the waiting time distribution, in two different ways: (i) in Section 4 we apply a powerful lemma of Bingham and Doney [4] for Laplace-Stieltjes transforms (LST) of regularly varying distributions to an expression of Le Gall [22] for the waiting time LST in the $M/G/1$ queue with ROS; (ii) in Section 5 we work out the general tail asymptotics of Section 3 for this case. Either way, the waiting time tail is proven to

exhibit the following behavior in the regularly varying case:

$$\mathsf{P}(W_{\rm ROS} > x) \sim \frac{\rho}{1-\rho} h(\rho) \mathsf{P}(B^{fw} > x), \quad x \to \infty. \tag{1.1}$$

Here, and throughout the paper, $f(x) \sim g(x)$ denotes $\lim_{x\to\infty} f(x)/g(x) = 1$; $h(\rho)$ is specified in Formulas (4.11) and (4.12). B^{fw} denotes the forward recurrence time of the service times, i.e., the residual service time. It is well-known that, with B denoting an arbitrary service time,

$$\mathsf{P}(B^{fw} > x) = \int_x^\infty \frac{\mathsf{P}(B > u)}{\mathsf{E}B} \mathrm{d}u, \quad x \geq 0. \tag{1.2}$$

Note that, except for Poisson arrivals, B^{fw} has a different distribution than the residual service requirement of the customer in service at arrival epochs.

Formula (1.1) should be compared with the waiting time tail asymptotics in the $M/G/1$ FCFS case [24]:

$$\mathsf{P}(W_{\rm FCFS} > x) \sim \frac{\rho}{1-\rho} \mathsf{P}(B^{fw} > x), \quad x \to \infty. \tag{1.3}$$

We shall show that $h(\rho) \leq 1$, which implies that ROS yields a (slightly) lighter tail than FCFS.

In Section 6 we allow the service time distribution to be completely general. We study the waiting time distribution in the case of heavy traffic (traffic load $\rho \uparrow 1$). When the service time variance is finite, we retrieve a result of Kingman [20]. When the service time distribution is regularly varying with infinite variance, we exploit a result of [8] to derive a new heavy-traffic limit theorem.

The paper ends with four appendices. Appendix A discusses several classes of heavy-tailed distributions. Appendices B and C provide the proofs of two theorems. In Appendix D we state and prove a lemma that is not explicitly used in the paper. However, it has been very useful in guiding us to the proofs of our main results. Essentially, the lemma states that when interested in events involving a large service time, we may in fact ignore the randomness in the arrival process.

Remark 1.1. A different way of randomly choosing a customer for service is the following. Put an arriving customer, who finds n waiting customers, with probability $\frac{1}{n+1}$ in one of the positions $1, 2, \ldots, n+1$, and serve customers according to their order in the queue. Fuhrmann and Iliadis [17] prove that this discipline gives rise to exactly the same waiting time distribution as ROS.

2 Preliminaries: Busy period and waiting time

We first focus on the busy period of the $GI/G/1$ queue. For the time being we may take the service discipline to be the familiar FCFS, since the busy period is the same for any non-idling discipline. At the end of this section – in Corollary 2.5 – we use the results on the busy period to state a useful relation for the waiting time in any non-idling service discipline.

Let us introduce some notation. The mean inter-arrival time is denoted with α and the random variable B stands for a generic service time, with mean $\mathsf{E}B = \beta$. A generic busy period is denoted with the random variable Z and τ is the number of customers served in a busy period. The residual busy period *as seen by an arriving customer* (i.e., the Palm version

associated with arrivals) is denoted with Z^{rp}. As before, we use B^{fw} to denote a random variable with the forward recurrence time distribution of the service times.

For the $GI/G/1$ queue the proof of the following proposition is given in [16]. The definitions of the widely used classes $\mathcal{S}^*$, $\mathcal{IRV}$, $\mathcal{L}$ and $\mathcal{D}$ can be found in Appendix A. The first proposition can be specialized to the $M/G/1$ queue by substituting $\mathsf{E}\tau = \frac{1}{1-\rho}$.

Proposition 2.1. *If $B \in \mathcal{S}^*$, then, for any $0 < c_1 < 1 < c_2$,*

$$\limsup_{n\to\infty} \frac{\mathsf{P}(\tau > n)}{\mathsf{E}\tau\mathsf{P}(B > c_1 n\alpha(1-\rho))} \leq 1, \tag{2.1}$$

and

$$\liminf_{n\to\infty} \frac{\mathsf{P}(\tau > n)}{\mathsf{E}\tau\mathsf{P}(B > c_2 n\alpha(1-\rho))} \geq 1. \tag{2.2}$$

Similarly, for any $0 < d_1 < 1 < d_2$,

$$\limsup_{x\to\infty} \frac{\mathsf{P}(Z > x)}{\mathsf{E}\tau\mathsf{P}(B > d_1 x(1-\rho))} \leq 1, \tag{2.3}$$

and

$$\liminf_{x\to\infty} \frac{\mathsf{P}(Z > x)}{\mathsf{E}\tau\mathsf{P}(B > d_2 x(1-\rho))} \geq 1. \tag{2.4}$$

In particular, if $B \in \mathcal{IRV}$ then

$$\mathsf{P}(\tau > n) \sim \mathsf{E}\tau\mathsf{P}(B > n\alpha(1-\rho)), \quad as \quad n \to \infty, \tag{2.5}$$

and

$$\mathsf{P}(Z > x) \sim \mathsf{E}\tau\mathsf{P}(B > x(1-\rho)), \quad as \quad x \to \infty. \tag{2.6}$$

The next proposition gives the asymptotics of the distribution of the *residual* busy period. Heuristically speaking, it indicates that a large residual busy period requires exactly one large service requirement (in the past). When analyzing waiting times (and residual service requirements) this result proves to be very useful as we shall see later. In fact, we shall sharpen the statement of the proposition (in line with the heuristics) in Corollary 2.3.

Proposition 2.2. *If $B \in \mathcal{L} \bigcap \mathcal{D}$, then*

$$\mathsf{P}(Z^{rp} > x) \sim \sum_{m=1}^{\infty} \mathsf{P}(B_{-m} > (x + m\alpha)(1-\rho)) \tag{2.7}$$

$$\sim \frac{\rho}{1-\rho}\mathsf{P}(B^{fw} > x(1-\rho)), \qquad x \to \infty, \tag{2.8}$$

where B_{-m} is the service time of the m-th customer (counting backwards) in the elapsed busy period.

Proof. Let us concentrate on the residual busy period as seen by an arbitrary customer ("customer 0") arriving at time $T_0 = 0$. With V_{-m} we denote the amount of work in the system found by customer $-m$ and by Z_{-m} the consecutive busy period if $V_{-m} = 0$. Furthermore,

for $m > 0$, T_{-m} is the time between the arrival of customer $-m$ and time 0 and T_m is the time of arrival of the m-th customer after time 0. We may write

$$\begin{aligned}
\mathsf{P}\left(Z^{rp} > x\right) &= \sum_{m=1}^{\infty} \mathsf{P}\left(V_{-m} = 0, Z_{-m} > T_{-m} + x\right) \\
&= \sum_{m=1}^{\infty} \mathsf{P}\left(V_{-m} = 0\right) \mathsf{P}\left(Z_{-m} > T_{-m} + x\right) \\
&= \frac{1}{\mathsf{E}\tau} \sum_{m=1}^{\infty} \mathsf{P}\left(Z_0 > T_m + x\right).
\end{aligned}$$

From this, the proof is quite straightforward in the case of constant inter-arrival times $T_m \equiv m\alpha$. In that case it follows from Proposition 2.1, that for any $\delta \in (0,1)$ and $d_1 > 1$ there is an x_0 such that

$$\mathsf{P}(Z_0 > T_m + x) \equiv \mathsf{P}(Z > m\alpha + x) \leq (1+\delta)\mathsf{E}\tau\mathsf{P}\left(B > d_1(x + m\alpha)(1-\rho)\right),$$

for all $x > x_0$ and $m \geq 1$. For $x > x_0$ this gives

$$\begin{aligned}
\mathsf{P}\left(Z^{rp} > x\right) &\leq (1+\delta) \sum_{m=1}^{\infty} \mathsf{P}\left(B > d_1(x + m\alpha)(1-\rho)\right) \\
&\sim \frac{(1+\delta)\rho}{d_1(1-\rho)} \mathsf{P}\left(B^{fw} > x(1-\rho)\right).
\end{aligned}$$

Now let $\delta \to 0$, $d_1 \to 1$ and use that $B^{fw} \in \mathcal{IRV}$ (by Property (7e) in Appendix A) to obtain the desired upper bound

$$\mathsf{P}\left(Z^{rp} > x\right) \leq (1+o(1))\frac{\rho}{1-\rho}\mathsf{P}\left(B^{fw} > x(1-\rho)\right), \qquad x \to \infty.$$

The lower bound can be derived similarly.

When inter-arrival times are not constant the proof is more involved since Z_0 and T_m are not independent. First we note that since $B \in \mathcal{L}$, then, for any $\varepsilon > 0$,

$$\mathrm{e}^{-\varepsilon x} = o\left(\mathsf{P}\left(B > x\right)\right), \qquad x \to \infty. \tag{2.9}$$

We shall now develop upper and lower bounds for $\sum_{m=0}^{\infty} \mathsf{P}\left(Z_0 > T_m + x\right)$, which coincide for $x \to \infty$.

Upper Bound. For any $\varepsilon \in (0,1)$,

$$\mathsf{P}(Z_0 > T_m + x) \leq \mathsf{P}\left(Z_0 > -\varepsilon x + m\alpha(1-\varepsilon) + x\right) + \mathsf{P}(T_m \leq -\varepsilon x + m\alpha(1-\varepsilon)).$$

From Proposition 2.1, for any $d_1 \in (0,1)$,

$$\sum_{m=0}^{\infty} \mathsf{P}(Z_0 > x(1-\varepsilon) + m\alpha(1-\varepsilon)) \leq (1+o(1))\mathsf{E}\tau \sum_{m=0}^{\infty} \mathsf{P}(B > d_1(1-\varepsilon)(1-\rho)(x+m\alpha)),$$

as $x \to \infty$. For notational convenience we set $c_1 = d_1(1-\varepsilon)$ and note that $c_1 \uparrow 1$ when $d_1 \uparrow 1$ and $\varepsilon \downarrow 0$. Furthermore,

$$\sum_{m=0}^{\infty} \mathsf{P}(B > c_1(1-\rho)(x+m\alpha)) \sim \frac{\rho}{c_1(1-\rho)}\mathsf{P}(B^{fw} > c_1(1-\rho)x).$$

We use t_n to denote the inter-arrival time of customer n and customer $n+1$, thus, $T_m = t_1 + \cdots + t_m$. By the Chernoff inequality we have, for any $r > 0$,

$$\mathsf{P}(T_m \le -\varepsilon x + m\alpha(1-\varepsilon)) = \mathsf{P}\left(\mathrm{e}^{-rT_m} \ge \mathrm{e}^{r\varepsilon x - rm\alpha(1-\varepsilon)}\right) \le \left(\mathsf{E}\mathrm{e}^{-rt_1}\right)^m \mathrm{e}^{rm\alpha(1-\varepsilon)-r\varepsilon x}.$$

Since $\mathsf{E}t_1 = \alpha$ and $\varepsilon > 0$, we can choose $r > 0$ sufficiently small, such that

$$\mathrm{e}^{r\alpha(1-\varepsilon)}\mathsf{E}\mathrm{e}^{-rt_1} < 1. \tag{2.10}$$

Then

$$\sum_{m=0}^{\infty} \mathsf{P}(T_m \le -\varepsilon x + m\alpha(1-\varepsilon)) \le \mathrm{e}^{-r\varepsilon x} \sum_{m=0}^{\infty} \left(\mathrm{e}^{r\alpha(1-\varepsilon)}\mathsf{E}\mathrm{e}^{-rt_1}\right)^m = \frac{\mathrm{e}^{-r\varepsilon x}}{1 - \mathrm{e}^{r\alpha(1-\varepsilon)}\mathsf{E}\mathrm{e}^{-rt_1}}. \tag{2.11}$$

Thus, from (2.9),

$$\limsup_{x\to\infty} \frac{\sum_{m=0}^{\infty} \mathsf{P}(Z_0 > T_m + x)}{\mathsf{P}(B^{fw} > c_1(1-\rho)x)} \le \mathsf{E}\tau \frac{\rho}{c_1(1-\rho)}.$$

Since $B^{fw} \in \mathcal{IRV}$ (Property (7e) in Appendix A), letting c_1 to 1, we get

$$\limsup_{x\to\infty} \frac{\sum_{m=0}^{\infty} \mathsf{P}(Z_0 > T_m + x)}{\mathsf{P}(B^{fw} > (1-\rho)x)} \le \mathsf{E}\tau \frac{\rho}{1-\rho},$$

which concludes the upper bound.

Lower Bound. For any $\varepsilon \in (0,1)$, put

$$n_{x,m} = \left\lfloor \frac{x(1+\varepsilon)}{\alpha} + \varepsilon m \right\rfloor,$$

where $\lfloor y \rfloor$ denotes the integer part of a positive real number y. Obviously,

$$\begin{aligned} \mathsf{P}(Z_0 > T_m + x) &\ge \mathsf{P}(Z_0 > T_{m+n_{x,m}}, T_{m+n_{x,m}} - T_m \ge x) \\ &\ge \mathsf{P}(Z_0 > T_{m+n_{x,m}}) - \mathsf{P}(T_{m+n_{x,m}} - T_m < x) \\ &= \mathsf{P}(\tau > m + n_{x,m}) - \mathsf{P}(T_{n_{x,m}} < x). \end{aligned}$$

From Proposition 2.1, for any $c_2 > 1$,

$$\mathsf{P}(\tau > m + n_{x,m}) \ge (1 + o(1))\mathsf{E}\tau\mathsf{P}(B > \alpha(1-\rho)(m + n_{x,m})c_2).$$

Similar to (2.10), we can choose $r > 0$ such that

$$\mathrm{e}^{r\alpha(1+\frac{1}{2}\varepsilon)/(1+\varepsilon)}\mathsf{E}\mathrm{e}^{-rt_1} < 1,$$

so that

$$\begin{aligned} \sum_{m=0}^{\infty} \mathsf{P}(T_{n_{x,m}} < x) &= \sum_{m=0}^{\infty} \mathsf{P}(\mathrm{e}^{-rT_{n_{x,m}}} > \mathrm{e}^{-rx}) \le \sum_{m=0}^{\infty} \mathrm{e}^{rx}\mathsf{E}\mathrm{e}^{-rT_{n_{x,m}}} = \sum_{m=0}^{\infty} \mathrm{e}^{rx}\left(\mathsf{E}\mathrm{e}^{-rt_1}\right)^{n_{x,m}} \\ &\le \frac{\mathrm{e}^{rx}\left(\mathsf{E}\mathrm{e}^{-rt_1}\right)^{x(1+\varepsilon)/\alpha - 1}}{1 - \left(\mathsf{E}\mathrm{e}^{-rt_1}\right)^{\varepsilon}} \le \frac{\mathrm{e}^{-\frac{1}{2}\varepsilon rx}}{\left(1 - \left(\mathsf{E}\mathrm{e}^{-rt_1}\right)^{\varepsilon}\right)\mathsf{E}e^{-rt_1}}. \end{aligned}$$

Thus, by (2.9),

$$\sum_{m=0}^{\infty} \mathsf{P}(Z_0 > T_m + x) \geq (1+o(1)) \frac{\rho \mathsf{E}\tau}{c_2(1+\varepsilon)(1-\rho)} \mathsf{P}(B^{fw} > (1-\rho)x(1+\varepsilon)c_2).$$

Letting $c_2 \downarrow 1$ and $\varepsilon \downarrow 0$, we get the desired result.

□

Remark 2.1. For the class of RV tails the equivalence (2.6) was proved by Zwart [29]. The asymptotic result (2.6) also holds in a class of so-called square-root insensitive subexponential distributions under the additional condition that the second moment of the inter-arrival time distribution is finite [18]. More precisely, Jelenkovic et al. [18] established the following result for the stable $G1/G1/1$ queue. If the following three conditions are satisfied:

(a) The distribution of service times is square-root insensitive:

$$\mathbf{P}(B > x + \sqrt{x}) \sim \mathbf{P}(B > x), \quad x \to \infty;$$

(b) also, the distribution of B belongs to the class of so-called *strong concave* (SC) distributions – which is a sub-class of $\mathcal{S}^*$;

(c) the distribution of inter-arrival times has a finite second moment;

then (2.6) holds. Under the same conditions, it may be shown that the distribution tail of the number of customers served in a busy period, τ, has similar asymptotics:

$$\mathbf{P}(\tau > n) \sim \mathbf{E}\tau \mathbf{P}(B > n\alpha(1-\rho)).$$

Therefore, one can conclude that the asymptotics (2.7)-(2.8) are also valid under conditions (a)-(c) above. It would be worthwhile to formulate and prove Corollaries 2.3 and 2.5 and Theorems 2.4 and 3.2 (the main theorem) for the class of service time distributions that satisfy (a) and (b) above, under the restriction that the arrival times satisfy (c).

Proposition 2.2 states that, for large x, the events $\bigcup_{m=1}^{\infty}\{B_{-m} > (x+m\alpha)(1-\rho)\}$ and $\{Z^{rp} > x\}$ are equally likely. In the sequel we shall need that these two events actually occur simultaneously (for large x). This statement is made precise in the following corollary.

Corollary 2.3. *If $B \in \mathcal{L} \bigcap \mathcal{D}$, then*

$$\mathsf{P}(Z^{rp} > x) \sim \sum_{m=1}^{\infty} \mathsf{P}(Z^{rp} > x, B_{-m} > (x+m\alpha)(1-\rho)) \tag{2.12}$$

$$\sim \mathsf{P}(Z^{rp} > x, \bigcup_{m=1}^{\infty} \{B_{-m} > (x+m\alpha)(1-\rho)\}), \qquad x \to \infty. \tag{2.13}$$

Proof. First we show that (2.12) implies (2.13). Note that

$$\begin{aligned}
&\mathsf{P}(Z^{rp} > x, \bigcup_{m=1}^{\infty} \{B_{-m} > (x + m\alpha)(1-\rho)\}) \\
&\leq \sum_{m=1}^{\infty} \mathsf{P}(Z^{rp} > x, B_{-m} > (x + m\alpha)(1-\rho)) \\
&\leq \mathsf{P}(Z^{rp} > x, \bigcup_{m=1}^{\infty} \{B_{-m} > (x + m\alpha)(1-\rho)\}) \\
&+ \sum_{m_1 \neq m_2} \mathsf{P}(B_{-m_1} > (x + m_1\alpha)(1-\rho), B_{-m_2} > (x + m_2\alpha)(1-\rho)),
\end{aligned}$$

where the last sum is not bigger than

$$\begin{aligned}
&\sum_{m_1=1}^{\infty} \sum_{m_2=1}^{\infty} \mathsf{P}(B_{-m_1} > (x + m_1\alpha)(1-\rho)) \mathsf{P}(B_{-m_2} > (x + m_2\alpha)(1-\rho)) \\
\sim \quad &\Big(\rho \mathsf{P}(B^{fw} > x)\Big)^2 = o(\mathsf{P}(B^{fw} > x)).
\end{aligned}$$

Using Proposition 2.2 this proves that (2.12) implies (2.13).

It remains to show that the right-hand side of (2.12) matches (2.8). As before, we use t_n to denote the inter-arrival time of customer n and customer $n+1$. Assume that for some constants $\varepsilon > 0$ and $R > 0$ the following events occur

1. $E_{m,1}^{\varepsilon,R}(x) := \Big\{B_{-m} > (x+R)\frac{1-\rho+\varepsilon(1+\rho)}{1-\varepsilon} + m\alpha(1-\rho) + \varepsilon m\alpha(1+\rho) + (1+\varepsilon)\alpha + 2R\Big\}$;
2. $E_{m,2}^{\varepsilon,R} := \{\text{for all } n \geq 1 : \sum_{i=1}^{n} B_{-m+i} \geq n\beta(1-\varepsilon) - R\}$;
3. $E_{m,3}^{\varepsilon,R} := \{\text{for all } n \geq 1 : \sum_{i=1}^{n} t_{-m+i} \leq n\alpha(1+\varepsilon) + R\}$;
4. $E_{4}^{\varepsilon,R} := \{\text{for all } n \geq 1 : \sum_{i=1}^{n} t_{i} \geq n\alpha(1-\varepsilon) - R\}$;

then V_n – the amount of work seen upon arrival by customer n – satisfies, for $n > -m$,

$$V_n \geq \sum_{i=-m}^{n-1} (B_i - t_{i+1}) \geq (x+R)\frac{1-\rho+\varepsilon(1+\rho)}{1-\varepsilon} - (n-1)\alpha(1-\rho+\varepsilon(1+\rho)).$$

Therefore all customers n with

$$n - 1 < \frac{x+R}{\alpha(1-\varepsilon)},$$

are in the same busy period, so that

$$Z^{rp} > \sum_{i=1}^{n} t_i \geq x.$$

We thus have

$$\begin{aligned}
\sum_{m=1}^{\infty} \mathsf{P}\left(E_{m,1}^{\varepsilon,R}(x) \cap E_{m,2}^{\varepsilon,R} \cap E_{m,3}^{\varepsilon,R} \cap E_{4}^{\varepsilon,R}\right) &\leq \sum_{m=1}^{\infty} \mathsf{P}\left(Z^{rp} > x, B > (x + m\alpha)(1-\rho)\right) \\
&\leq \sum_{m=1}^{\infty} \mathsf{P}\left(B > (x + m\alpha)(1-\rho)\right),
\end{aligned}$$

and we are done if a lower bound for $\sum_{m=1}^{\infty} \mathsf{P}\left(E_{m,1}^{\varepsilon,R}(x) \cap E_{m,2}^{\varepsilon,R} \cap E_{m,3}^{\varepsilon,R} \cap E_4^{\varepsilon,R}\right)$ is shown to be arbitrarily close to $(1+o(1))\frac{\rho}{1-\rho}\mathsf{P}\left(B^{fw} > x(1-\rho)\right)$, as $x \to \infty$. For any fixed $\delta > 0$ we can find (by the strong law of large numbers) ε and R such that $\mathsf{P}\left(E_{m,2}^{\varepsilon,R}\right) \geq 1-\delta$ and $\mathsf{P}\left(E_{m,3}^{\varepsilon,R} \cap E_4^{\varepsilon,R}\right) \geq 1-\delta$. Thus, as $x \to \infty$,

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(E_{m,1}^{\varepsilon,R}(x) \cap E_{m,2}^{\varepsilon,R} \cap E_{m,3}^{\varepsilon,R} \cap E_4^{\varepsilon,R}\right) = \sum_{m=1}^{\infty} \mathsf{P}\left(E_{m,1}^{\varepsilon,R}(x)\right) \mathsf{P}\left(E_{m,2}^{\varepsilon,R}\right) \mathsf{P}\left(E_{m,3}^{\varepsilon,R} \cap E_4^{\varepsilon,R}\right) \\
&\geq (1-\delta)^2 \sum_{m=1}^{\infty} \mathsf{P}\left(E_{m,1}^{\varepsilon,R}(x)\right) \\
&\sim \frac{(1-\delta)^2\rho}{1-\rho+\varepsilon(1+\rho)} \mathsf{P}\left(B^{fw} > (x+R)\frac{1-\rho+\varepsilon(1+\rho)}{1-\varepsilon} + (1+\varepsilon)\alpha + 2R\right) \\
&\sim \frac{(1-\delta)^2\rho}{1-\rho+\varepsilon(1+\rho)} \mathsf{P}\left(B^{fw} > x\frac{1-\rho+\varepsilon(1+\rho)}{1-\varepsilon}\right),
\end{aligned}$$

where we have used $B^{fw} \in \mathcal{L}$. Now, first letting $\varepsilon \to 0$, using $B^{fw} \in \mathcal{IRV}$ (Property (7e) in Appendix A), and then $\delta \to 0$ the proof is completed. □

Remark 2.2. Expression (2.12) shows that the occurrence of a large residual busy period is due to a single large service time *in the past*. This can be explained as follows. The busy period is the sum of services of the customers in that busy period. The number of customers in the busy period after time 0 (the point of arrival) is almost surely finite. There are, however, infinitely many service times in the past, each of them being potentially large. This leads to the integrated tail of the service time distribution.

Besides the busy period and the residual busy period, there is a third entity whose distribution is the same for all non-idling and non-preemptive service disciplines: B^{rp}, the residual service requirement of the customer in service (if any) upon arrival of a new customer. The tail asymptotics for the distribution of B^{rp} are determined in the following theorem. Not only is the theorem of interest in itself, but several steps in its proof will also be useful in proving our main result in Theorem 3.2.

Theorem 2.4. *If $B \in \mathcal{L} \bigcap \mathcal{D}$ then, for any non-preemptive and non-idling service discipline,*

$$\mathsf{P}\left(B^{rp} > x\right) \sim \rho \mathsf{P}\left(B^{fw} > x\right),$$

as $x \to \infty$.

Proof. See Appendix B. □

Corollary 2.5. *If $B \in \mathcal{L} \bigcap \mathcal{D}$, then for any non-preemptive and non-idling service discipline, the waiting time W and residual busy period Z^{rp} seen by a customer arriving to a stationary $GI/G/1$ queue satisfy $W \leq Z^{rp}$ a.s. and therefore, as $x \to \infty$,*

$$\mathsf{P}(W > x) \sim \sum_{m=1}^{\infty} \mathsf{P}\left(W > x, B_{-m} > (x+m\alpha)(1-\rho)\right), \qquad x \to \infty. \tag{2.14}$$

Proof. Since the service discipline is non-preemptive we have $W \geq B^{rp}$ almost surely, so that by Theorem 2.4,

$$\mathsf{P}\left(W > x\right) \geq \mathsf{P}\left(B^{rp} > x\right) \sim \rho \mathsf{P}\left(B^{fw} > x\right), \qquad x \to \infty.$$

Thus, in the following, we may neglect terms which are $o\left(\mathsf{P}\left(B^{fw} > x\right)\right)$. Using $W \leq Z^{rp}$ (almost surely) and Corollary 2.3 we therefore have (similar to the proof of Theorem 2.4)

$$\begin{aligned}
\mathsf{P}(W > x) &= \mathsf{P}\left(W > x, Z^{rp} > x\right) \\
&\sim \mathsf{P}(W > x, Z^{rp} > x, \bigcup_{m=1}^{\infty} \{B_{-m} > (x + m\alpha)(1-\rho)\}) \\
&\sim \sum_{m=1}^{\infty} \mathsf{P}(W > x, Z^{rp} > x, B_{-m} > (x + m\alpha)(1-\rho)) \\
&= \sum_{m=1}^{\infty} \mathsf{P}(W > x, B_{-m} > (x + m\alpha)(1-\rho)).
\end{aligned}$$

□

3 Random Order of Service

We now turn to the $GI/GI/1$ queue with Random Order of Service. We start with analyzing the waiting time *conditional* on the initial queue length q, none of these customers having received previous service. It is convenient to associate service times with customers in their order of service instead of their order of arrival. The customer which is served first has a service time B_1, the second has B_2, etc. Denote with $W_{\mathrm{ROS}}(q)$ the conditional waiting time of an arbitrary customer in the queue.

The following lemma does not require any assumptions on the distributions of service times and inter-arrival times. It shows that $W_{\mathrm{ROS}}(q)/q$ converges in distribution to a random variable whose distribution has support $[0, \frac{\beta}{1-\rho}]$. Note that this contrasts with the FCFS queue, in which case the corresponding quantity $W(q)/q$ (for the last customer in line) converges to the constant β almost surely.

Lemma 3.1. *As $q \to \infty$,*

$$\mathsf{P}\left(W_{\mathrm{ROS}}(q) > \frac{\beta q}{1-\rho} y\right) \to \left((1-y)^{+}\right)^{\frac{1}{1-\rho}}, \tag{3.1}$$

uniformly in $y \in [0, \infty)$.

Proof. Since the limiting distribution is continuous and non-defective, by the monotonicity of probability distribution functions, it is sufficient to prove point-wise convergence.

Let us thus fix $y \in (0,1)$. For $n = 1, 2, \ldots$, denote by Q_n the number of customers in the queue at the time instant of the nth service completion. We define the event A_1 by

$$A_1 := \left\{ Q_i \in [q(1-\varepsilon) - i(1-\rho+\varepsilon), q(1+\varepsilon) - i(1-\rho-\varepsilon)] \quad \forall\, i = 1, 2, \ldots, \frac{q}{1-\rho} \right\}.$$

By the strong law of large numbers, for any $\varepsilon > 0$, there exists $\tilde{q} \equiv \tilde{q}(\varepsilon)$ such that $\mathsf{P}(A_1) \geq 1-\varepsilon$ for all $q \geq \tilde{q}$.

Let $q_i^- = q(1-\varepsilon) - i(1-\rho+\varepsilon)$ and $q_i^+ = q(1+\varepsilon) - i(1-\rho-\varepsilon)$ and denote $N(v) = \min\{n : \sum_1^n B_i > v\}$; customer $N(v)$ is in service at time v. Defining

$$A_2 := \left\{ N(v) \in \left[\frac{v}{\beta}(1-\varepsilon) - R, \frac{v}{\beta}(1+\varepsilon) + R \right], \quad \forall\, v \in \left[0, \frac{\beta q}{1-\rho} y\right] \right\},$$

for any $\varepsilon > 0$, we may choose $R > 0$ such that $\mathsf{P}(A_2) \geq 1-\varepsilon$. Thus, $\mathsf{P}(A_1 \bigcap A_2) \geq 1-2\varepsilon$ and

$$\mathsf{P}\left(W_{\mathrm{ROS}}(q) > \frac{\beta q}{1-\rho} y \right) = P(y) + O(\varepsilon),$$

where

$$P(y) := \mathsf{P}\left(W_{\mathrm{ROS}}(q) > \frac{\beta q}{1-\rho} y, A_1 \bigcap A_2 \right).$$

We further define $u = \frac{\beta q}{1-\rho} y$, $n^-(u) = \frac{u}{\beta}(1-\varepsilon) - R$ and $n^+(u) = \frac{u}{\beta}(1+\varepsilon) + R$. Since $\{W_{\mathrm{ROS}}(q) > \frac{\beta q}{1-\rho} y\}$ implies that customer 0 was not selected in the first $N\left(\frac{\beta q}{1-\rho} y\right)$ trials, we have, as q and u tend to infinity keeping y constant,

$$\begin{aligned}
P(y) &\leq \Pi_{i=1}^{n^-(u)} \left(1 - \frac{1}{q_i^+}\right) \\
&= (1+o(1)) \exp\left(- \sum_{i=1}^{n^-(u)} \frac{1}{q_i^+} \right) \\
&= (1+o(1)) \exp\left(- \int_0^{n^-(u)} \frac{1}{q(1+\varepsilon) - v(1-\rho-\varepsilon)} dv \right) \\
&= (1+o(1)) \left(1 - y \frac{(1-\varepsilon)(1-\rho-\varepsilon)}{(1+\varepsilon)(1-\rho)} \right)^{\frac{1}{1-\rho-\varepsilon}} \\
&= (1+o(1))(1 - y + O(\varepsilon))^{\frac{1}{1-\rho-\varepsilon}}.
\end{aligned}$$

Similarly,

$$P(y) \geq \Pi_{i=1}^{n^+(u)} \left(1 - \frac{1}{q_i^-}\right) - O(\varepsilon) = (1+o(1)) \left(1 - y - O(\varepsilon)\right)^{\frac{1}{1-\rho-\varepsilon}} - O(\varepsilon).$$

Letting ε pass to 0, we obtain (3.1). □

The main result of our paper is stated in the next theorem.

Theorem 3.2. *In the GI/G/1 ROS queue with $B \in \mathcal{L} \bigcap \mathcal{D}$, we have*

$$\begin{aligned}
\mathsf{P}(W_{\mathrm{ROS}} > x) &\sim \rho \mathsf{P}(B^{fw} > x) \\
&+ \int_0^{cx} dv \int_{(v\alpha+x)(1-\rho)}^{v\alpha+x} d\mathsf{P}(B \leq z) \left(1 - \frac{(x+v\alpha-z)(1-\rho)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&+ \int_{cx}^{\infty} dv \int_{v\alpha}^{v\alpha+x} d\mathsf{P}(B \leq z) \left(1 - \frac{(x+v\alpha-z)(1-\rho)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&+ \int_{cx}^{\infty} dv \int_{(v\alpha+x)(1-\rho)}^{v\alpha} d\mathsf{P}(B \leq z) \left(1 - \frac{x(1-\rho)}{z - v\alpha(1-\rho)}\right)^{\frac{1}{1-\rho}},
\end{aligned}$$

where $c = \frac{1-\rho}{\alpha\rho}$.

Proof. In Appendix C. □

Letting $W^*_{\rm ROS}$ be a random variable independent of B with the limiting distribution of $W_{\rm ROS}(q)/q$, as $q \to \infty$, we may conveniently rewrite the above as:

$$\begin{aligned}
\mathsf{P}(W_{\rm ROS} > x) &\sim \mathsf{P}(B^{rp} > x) \\
&+ \int_0^{cx} dv \mathsf{P}\left((v\alpha + x)(1-\rho) < B \le v\alpha + x, W^*_{\rm ROS} > \frac{\alpha(x + v\alpha - B)}{B}\right) \\
&+ \int_{cx}^{\infty} dv \mathsf{P}\left(v\alpha < B \le v\alpha + x, W^*_{\rm ROS} > \frac{\alpha(x + v\alpha - B)}{B}\right) \\
&+ \int_{cx}^{\infty} dv \mathsf{P}\left((v\alpha + x)(1-\rho) < B \le v\alpha, W^*_{\rm ROS} > \frac{\beta x}{B - v\alpha(1-\rho)}\right).
\end{aligned}$$

This allows for the following interpretation: The waiting time is larger than x when one of the following occurs:

1. (first term) The customer in service has a residual service time larger than x. Recall that, by Theorem 2.4, $\mathsf{P}\left(B^{rp} > x\right) \sim \rho \mathsf{P}\left(B^{fw} > x\right)$.

2. (second term) A customer (with index $-v$) that arrived at some time $-t = -\alpha v$ between time $-\alpha c x$ and time 0 required a service z larger than $(t+x)(1-\rho)$ but smaller than $t+x$. The service times of other customers in the system at time $-t$ are negligible compared to z. The large service time ends at time $-t + z \in (0, x)$, leaving approximately z/α competing customers in the system. Customer 0 thus waits for approximately $-t + z + W^*_{\rm ROS} z/\alpha$. Thus $W^*_{\rm ROS}$ needs to be larger than $(x + t - z)\alpha/z$.

3. (third term) This term is similar to the previous. Now, the large customer arrived at time $-t < -\alpha c x$ with a service requirement $z \in (t, t + x)$. That customer thus leaves at time $-t + z \in (0, x)$ with approximately z/α customers in the system.

4. (fourth term) Again, the large customer arrived at time $-t < -\alpha c x$, but leaves before time 0: its service requirement is $z \in ((t + x)(1 - \rho), t)$. Neglecting the size of the customer in service at time 0, the "service lottery" starts immediately upon arrival of customer 0. The number of competing customers at time 0 is approximately $t/\alpha - (t - z)/\beta$ which is the number of arrivals minus the number of departures between times $-t$ and 0. Therefore, the waiting time of customer 0 is larger than x if $W^*_{\rm ROS}$ is larger than $x/(t/\alpha - (t - z)/\beta) = \beta x/(z - t(1 - \rho))$.

4 The $M/G/1$ queue with regularly varying service time distribution

In this section we restrict ourselves to the case of Poisson arrivals and regularly varying service time distribution. In this case we are able to obtain detailed tail asymptotics for the waiting time distribution by applying Lemma A.1 for Laplace-Stieltjes transforms (LST) of regularly varying distributions to an expression of Le Gall [22] for the waiting time LST in the $M/G/1$ queue with ROS. In Section 5, we shall present an alternative approach to the same result, viz., we shall work out the general tail asymptotics of Section 3 for this case.

We consider an $M/G/1$ queue with arrival rate $\lambda = 1/\alpha$ and service time distribution $B(\cdot)$ with mean β and LST $\beta\{\cdot\}$. As before, the load of the queue is $\rho \stackrel{\text{def}}{=} \lambda\beta < 1$.

The LST of the waiting time distribution for the ROS discipline is given by (see Le Gall [22] or Cohen [13], p. 439):

$$\mathsf{E}[\mathrm{e}^{-sW_{\mathrm{ROS}}}] = 1 - \rho + \rho\beta^{fw}\{s\} - \frac{\rho(1-\rho)}{\beta s}\int_{\mu\{s\}}^{1} \frac{\partial\Phi}{\partial z}(s,z)\Psi(s,z)\mathrm{d}z, \tag{4.1}$$

with

$$\Phi(s,z) \stackrel{\text{def}}{=} (1-z)\frac{\beta\{s+\lambda(1-z)\} - \beta\{\lambda(1-z)\}}{z - \beta\{\lambda(1-z)\}}, \tag{4.2}$$

$$\Psi(s,z) \stackrel{\text{def}}{=} \exp\left[-\int_z^1 \frac{\mathrm{d}y}{y - \beta\{s+\lambda(1-y)\}}\right], \tag{4.3}$$

where $\beta^{fw}\{\cdot\}$ is the LST of the forward recurrence time of the service time:

$$\beta^{fw}\{s\} \stackrel{\text{def}}{=} \frac{1-\beta\{s\}}{\beta s}, \tag{4.4}$$

and $\mu\{s\}$ is the LST of the busy period distribution, satisfying the relation

$$\mu\{s\} - \beta\big\{s + \lambda(1-\mu\{s\})\big\} = 0. \tag{4.5}$$

It is possible to rewrite (4.1) in a simpler form using integration by parts. Indeed

$$\int_{\mu\{s\}}^{1} \frac{\partial\Phi}{\partial z}(s,z)\Psi(s,z)\mathrm{d}z = \Big[\Phi(s,z)\Psi(s,z)\Big]_{\mu\{s\}}^{1} - \int_{\mu\{s\}}^{1} \frac{\Phi(s,z)\Psi(s,z)\mathrm{d}z}{z - \beta\{s+\lambda(1-z)\}},$$

and the simple relations

$$\Phi(s,\mu(s)) = 1 - \mu\{s\}, \quad \Phi(s,1) = \frac{1-\beta\{s\}}{1-\rho},$$

$$\Psi(s,\mu\{s\}) = 0, \quad \Psi(s,1) = 1,$$

yield

$$\Big[\Phi(s,z)\Psi(s,z)\Big]_{\mu\{s\}}^{1} = \frac{1-\beta\{s\}}{1-\rho}.$$

Using this relation and (4.4) in (4.1) yields the following simpler alternative:

$$\mathsf{E}[\mathrm{e}^{-sW_{\mathrm{ROS}}}] = 1 + \frac{\rho(1-\rho)}{\beta s}\int_{\mu\{s\}}^{1} \widehat{\Phi}(s,z)\Psi(s,z)\mathrm{d}z, \tag{4.6}$$

with

$$\begin{aligned}\widehat{\Phi}(s,z) &\stackrel{\text{def}}{=} \frac{\Phi(s,z) - \frac{\beta s}{1-\rho}}{z - \beta\{s+\lambda(1-z)\}} \\ &= \frac{1 - z - \frac{\beta s}{1-\rho}}{z - \beta\{s+\lambda(1-z)\}} - \frac{1-z}{z-\beta\{\lambda(1-z)\}}.\end{aligned}$$

As before, we write $f(x) \sim g(x)$ if $f(x)/g(x) \to 1$ when $x \to \infty$, and similarly we write $a(s) \sim b(s)$ if $a(s)/b(s) \to 1$ when $s \to 0$. In this section and the next we assume that the service requirement distribution $B(\cdot)$ is regularly varying with index $-\nu$, $1 < \nu < 2$:

$$\mathsf{P}(B > x) \sim \frac{C}{\Gamma(1-\nu)} x^{-\nu} L(x), \quad x \to \infty, \tag{4.7}$$

with C a constant, $\Gamma(\cdot)$ the Gamma function and $L(\cdot)$ a slowly varying function at infinity, cf. [5].

Lemma A.1 in Appendix A implies, in combination with the assumption that (4.7) holds with $\nu \in (1,2)$:

$$\beta\{s\} - 1 + \beta s \sim C s^{\nu} L(1/s), \text{ as } s \to 0. \tag{4.8}$$

In addition, cf. (4.4),

$$1 - \beta^{fw}\{s\} \sim \frac{C}{\beta} s^{\nu-1} L(1/s), \text{ as } s \to 0. \tag{4.9}$$

De Meyer and Teugels [23] have proven that the busy period distribution of an $M/G/1$ queue with regularly varying service time distribution is also regularly varying at infinity. More precisely:

$$\mu\{s\} - 1 + \frac{\beta s}{1-\rho} \sim \frac{C}{(1-\rho)^{\nu+1}} s^{\nu} L(1/s), \text{ as } s \to 0. \tag{4.10}$$

We shall use the first-order behavior of $\mu\{s\}$ further on.

Our goal is to prove the following theorem.

Theorem 4.1. *If the service time distribution in the $M/G/1$ queue operating under the ROS discipline is regularly varying at infinity with index $-\nu \in (-2,-1)$, then the waiting time distribution is regularly varying at infinity with index $1-\nu \in (-1,0)$. More precisely, if (4.7) holds then, as $x \to \infty$,*

$$\begin{aligned} \mathsf{P}(W_{\rm ROS} > x) &\sim \frac{\rho}{1-\rho} h(\rho,\nu) \mathsf{P}(B^{fw} > x) \\ &\sim \frac{\rho}{1-\rho} h(\rho,\nu) \frac{1}{\Gamma(2-\nu)} \frac{C}{\beta} x^{1-\nu} L(x), \end{aligned}$$

with

$$h(\rho,\nu) \stackrel{\rm def}{=} \int_0^1 f(u,\rho,\nu) \mathrm{d}u, \tag{4.11}$$

$$f(u,\rho,\nu) \stackrel{\rm def}{=} \frac{\rho}{1-\rho} \left(\frac{\rho u}{1-\rho} \right)^{\nu-1} (1-u)^{\frac{1}{1-\rho}} + \left(1 + \frac{\rho u}{1-\rho} \right)^{\nu} (1-u)^{\frac{1}{1-\rho}-1}. \tag{4.12}$$

Remark 4.1. The following result for the waiting time distribution in the $M/G/1$ queue operating under the FCFS discipline is well-known (cf. Cohen [12] for the regularly varying case; see Pakes [24] for an extension to the larger class of subexponential residual service time distributions): if (4.7) holds, then

$$\mathsf{P}(W_{\rm FCFS} > x) \sim \frac{\rho}{1-\rho} \mathsf{P}(B^{fw} > x), \quad x \to \infty. \tag{4.13}$$

We can now conclude that

$$\mathsf{P}(W_{\rm ROS} > x) \sim h(\rho,\nu) \mathsf{P}(W_{\rm FCFS} > x), \quad x \to \infty. \tag{4.14}$$

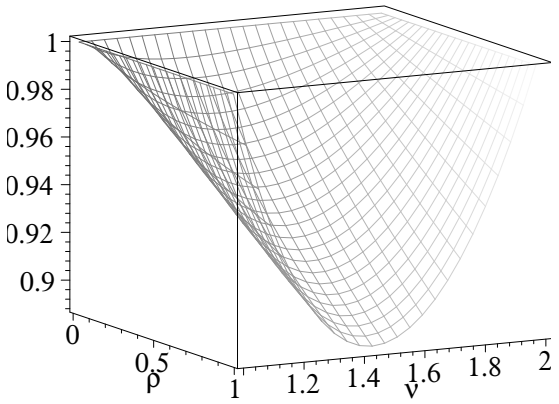

Figure 1: A plot of the function $(\rho, \nu) \mapsto h(\nu, \rho)$ for $0 \leq \rho \leq 1$ and $1 \leq \nu \leq 2$. Note that the minimal value for h seems to be $h(1, 3/2) = \Gamma(3/2) \approx 0.88622\cdots$.

Remark 4.2. It is actually possible to prove that $h(\rho, \nu)$ is always less than 1: indeed, the function $f(u, \rho, \nu)$ is strictly convex in ν (as a sum of exponentials) and therefore $h(\rho, \nu)$ is also strictly convex in ν; moreover, simple computations using integration by parts show that

$$\int_0^1 f(u, \rho, 1)\mathrm{d}u = \int_0^1 f(u, \rho, 2)\mathrm{d}u = 1,$$

and therefore

$$h(\rho, \nu) < 1, \quad \text{for all } 1 < \nu < 2,\ 0 \leq \rho < 1. \tag{4.15}$$

Numerical computations with MAPLE suggest that $h(\rho, \nu)$ is decreasing in ρ and thus always larger than its limit when $\rho \to 1$, which is by simple arguments equal to $\Gamma(\nu)$. Unfortunately, we have not been able to find a simple proof for this fact. In Figure 1 we have plotted $h(\rho, \nu)$ for $0 < \rho < 1$ and $1 < \nu < 2$.
It is interesting to observe that the tail behavior of W_{ROS} and W_{FCFS} is so similar in the regularly varying case. This strongly contrasts with the situation for the $M/M/1$ queue, where the purely exponential waiting time tail for FCFS strongly deviates from the $C_0 x^{-5/6}\mathrm{e}^{-C_1 x - C_2 x^{1/3}}$ waiting time tail behavior that was exposed by Flatto [15] for ROS.

Remark 4.3. The first part of $h(\rho, \nu)$ is a Beta function, and the second part is a hypergeometric function (cf. Abramowitz and Stegun [1]). In particular,

$$\int_0^1 u^{\nu-1}(1-u)^{\frac{1}{1-\rho}}\mathrm{d}u = B\Big(\nu, \frac{1}{1-\rho} + 1\Big) = \frac{\Gamma(\nu)\Gamma(\frac{1}{1-\rho}+1)}{\Gamma(\nu + \frac{1}{1-\rho}+1)}.$$

Using partial integration and the above formula for Beta functions, one gets the following form, which is useful for future comparisons (see Section 5):

$$h(\rho, \nu) = 1 - \rho + \int_0^1 g(u, \rho, \nu)\mathrm{d}u, \tag{4.16}$$

with

$$g(u, \rho, \nu) \stackrel{\text{def}}{=} \left[\frac{\rho\nu(1-u)-\rho}{u}\left(\frac{\rho u}{1-\rho}\right)^{\nu-1} + \rho\nu\left(1 + \frac{\rho u}{1-\rho}\right)^{\nu-1}\right](1-u)^{\frac{1}{1-\rho}}.$$

Proof of Theorem 4.1. We shall prove the theorem by applying Lemma A.1 to an expression for the LST of the waiting time distribution. So we need to consider $\mathsf{E}[\mathrm{e}^{-sW_{\mathrm{ROS}}}]-1$ as $s\to 0$. In order to do that, we use the change of variable $1-z=\varepsilon(s)u$, with $\varepsilon(s)\stackrel{\mathrm{def}}{=}1-\mu\{s\}$, in (4.6) to obtain

$$\int_{\mu\{s\}}^{1}\widehat{\Phi}(s,z)\Psi(s,z)\mathrm{d}z=\varepsilon(s)\int_0^1\widehat{\Phi}(s,1-\varepsilon(s)u)\Psi(s,1-\varepsilon(s)u)\mathrm{d}u.$$

Let us first evaluate the function Ψ. One can write

$$\log\Psi(s,1-\varepsilon(s)u)=-\int_0^1\frac{\varepsilon(s)u\mathrm{d}v}{1-\varepsilon(s)uv-\beta\{s+\lambda\varepsilon(s)uv\}},$$

and, using (4.8) and (4.10),

$$\lim_{s\to 0}\frac{1}{\varepsilon(s)}\Big[1-\varepsilon(s)uv-\beta\{s+\lambda\varepsilon(s)uv\}\Big]=(1-\rho)(1-uv),$$

the limit being uniform in u and v. Therefore the following limit holds uniformly in u:

$$\lim_{s\to 0}\Psi(s,1-\varepsilon(s)u)=\exp\left[-\frac{1}{1-\rho}\int_0^1\frac{u\mathrm{d}v}{1-uv}\right]=(1-u)^{\frac{1}{1-\rho}}. \tag{4.17}$$

The evaluation of $\widehat{\Phi}$ is not difficult either. First note that the denominators appearing in $\widehat{\Phi}(s,1-\varepsilon(s)u)$ can be expressed in terms of $\beta^{fw}\{\cdot\}$ as follows:

$$\begin{aligned}
&1-\varepsilon(s)u-\beta\{s+\lambda\varepsilon(s)u\}\ =\\
&\qquad -(1-\rho)\varepsilon(s)u+\beta s-(\beta s+\rho\varepsilon(s)u)(1-\beta^{fw}\{s+\lambda\varepsilon(s)u\}),\\
&1-\varepsilon(s)u-\beta\{\lambda\varepsilon(s)u\}\ =\\
&\qquad -(1-\rho)\varepsilon(s)u-\rho\varepsilon(s)u(1-\beta^{fw}\{\lambda\varepsilon(s)u\}).
\end{aligned}$$

Then write

$$\begin{aligned}
&\Big(\varepsilon(s)u-\frac{\beta s}{1-\rho}\Big)(1-\varepsilon(s)u-\beta\{\lambda\varepsilon(s)u\})-\varepsilon(s)u(1-\varepsilon(s)u-\beta\{s+\lambda\varepsilon(s)u\})\\
&\qquad =\ \varepsilon(s)u\Big[(\beta s+\rho\varepsilon(s)u)(1-\beta^{fw}\{s+\lambda\varepsilon(s)u\})\\
&\qquad\qquad +\Big(\frac{\rho\beta s}{1-\rho}-\rho\varepsilon(s)u\Big)(1-\beta^{fw}\{\lambda\varepsilon(s)u\})\Big],
\end{aligned}$$

and finally, as $s\to 0$,

$$\begin{aligned}
\widehat{\Phi}(s,1-\varepsilon(s)u)\ \sim\ &-\frac{1}{(1-\rho)(1-u)}\Big[\Big(1+\frac{\rho u}{1-\rho}\Big)(1-\beta^{fw}\{s+\lambda\varepsilon(s)u\})\\
&\qquad +\frac{\rho(1-u)}{1-\rho}(1-\beta^{fw}\{\lambda\varepsilon(s)u\})\Big].
\end{aligned}$$

We take into account now the regular variation assumption (4.7), which yields (4.9) so that

$$\begin{aligned}
\widehat{\Phi}(s,1-\varepsilon(s)u)\ \sim\ &-\frac{Cs^{\nu-1}}{\beta(1-\rho)(1-u)}\Big\{\frac{\rho(1-u)}{1-\rho}\Big(\frac{\rho u}{1-\rho}\Big)^{\nu-1}L\Big(\frac{1}{\lambda\varepsilon(s)u}\Big)\\
&\qquad +\Big(1+\frac{\rho u}{1-\rho}\Big)^{\nu}L\Big(\frac{1}{s+\lambda\varepsilon(s)u}\Big)\Big\}.
\end{aligned}$$

Using Potter's Theorem (see [23], Theorem 1.5.6), setting $\delta \stackrel{\text{def}}{=} (\nu-1)/2$, there exists $X > 0$ such that, as long as $s \le 1/X$ and $\lambda\varepsilon(s) \le 1/X$,

$$\begin{aligned} L\Big(\frac{1}{\lambda\varepsilon(s)u}\Big) &\le 2\max\Big[\Big(\frac{s}{\lambda\varepsilon(s)u}\Big)^{\delta}, \Big(\frac{s}{\lambda\varepsilon(s)u}\Big)^{-\delta}\Big] L(1/s), \\ L\Big(\frac{1}{s+\lambda\varepsilon(s)u}\Big) &\le 2\max\Big[\Big(\frac{1}{1+\frac{\lambda\varepsilon(s)}{s}u}\Big)^{\delta}, \Big(\frac{1}{1+\frac{\lambda\varepsilon(s)}{s}u}\Big)^{-\delta}\Big] L(1/s). \end{aligned}$$

These bounds allow application of the Dominated Convergence Theorem to

$$\int_0^1 \frac{1}{s^{\nu-1}L(1/s)} \widehat{\Phi}(s, 1-\varepsilon(s)u)\Psi(s, 1-\varepsilon(s)u)\mathrm{d}u,$$

which yields

$$1 - \mathsf{E}[\mathrm{e}^{-sW_{\mathrm{ROS}}}] \sim \frac{\lambda C}{1-\rho}\Big[\int_0^1 f(u,\rho,\nu)\mathrm{d}u\Big] s^{\nu-1}L(1/s), \text{ as } s \to 0. \tag{4.18}$$

Using Lemma A.1, the theorem follows. □

5 Agreement of results

While Sections 3 and 4 use completely different methods of proof, it is clear that the asymptotics for the $M/G/1$ queue with regularly varying service time distribution (as in Theorem 4.1) has to be a mere consequence of Theorem 3.2. This section shows that this is indeed true. As a first step, we give another asymptotic expression for $\mathsf{P}(W_{\mathrm{ROS}} > x)$ which, while less intuitive than the rewriting proposed in Section 3, bears a strong similarity with Theorem 4.1.

Lemma 5.1. *In the GI/G/1 ROS queue with $B \in \mathcal{L} \bigcap \mathcal{D}$, we have*

$$\begin{aligned} \mathsf{P}(W_{\mathrm{ROS}} > x) \sim\ & \rho\mathsf{P}(B^{fw} > x) \\ & + \int_0^1 \Big\{ \frac{1}{\alpha(1-\rho)} \mathsf{E}\Big[B\mathbb{1}_{\{\frac{x(1-\rho)}{\rho u+1-\rho} < B \le \frac{x(1-\rho)}{\rho u}\}}\Big] \\ & \qquad + \frac{x}{\alpha u^2}\mathsf{P}\Big(B > \frac{x(1-\rho)}{\rho u}\Big)(1-u)^{\frac{1}{1-\rho}}\,\mathrm{d}u \Big\}. \end{aligned}$$

Proof. To simplify the computations, assume that B has a density function b. This is, however, not really needed for the result.

The first term in Theorem 3.2 coincides with that in the expression above. Next, consider the second and third terms in Theorem 3.2 together, using the changes of variable $z \mapsto u =$

$(x+v\alpha-z)(1-\rho)/(\rho z)$ and $v \mapsto t = (x+v\alpha)(1-\rho)/(\rho u+1-\rho)$.

$$
\begin{aligned}
&\int_0^{cx} \mathrm{d}v \int_{(v\alpha+x)(1-\rho)}^{v\alpha+x} \mathrm{d}z\, b(z) \left(1 - \frac{(x+v\alpha-z)(1-\rho)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&\quad + \int_{cx}^{\infty} \mathrm{d}v \int_{v\alpha}^{v\alpha+x} \mathrm{d}z\, b(z) \left(1 - \frac{(x+v\alpha-z)(1-\rho)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&\quad = \int_0^{\infty} \mathrm{d}v \int_{(v\alpha+x)(1-\rho)}^{v\alpha+x} \mathrm{d}z\, b(z) \left(1 - \frac{(x+v\alpha-z)(1-\rho)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&\qquad - \int_{cx}^{\infty} \mathrm{d}v \int_{(v\alpha+x)(1-\rho)}^{v\alpha} \mathrm{d}z\, b(z) \left(1 - \frac{(x+v\alpha-z)(1-\rho)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&\quad = \int_0^{\infty} \mathrm{d}v \int_0^1 \mathrm{d}u\, b\left(\frac{(x+v\alpha)(1-\rho)}{\rho u+1-\rho}\right) \frac{\rho(x+v\alpha)(1-\rho)}{(\rho u+1-\rho)^2} (1-u)^{\frac{1}{1-\rho}} \\
&\qquad - \int_{cx}^{\infty} \mathrm{d}v \int_0^{\frac{cx}{v}} \mathrm{d}u\, b\left(\frac{(x+v\alpha)(1-\rho)}{\rho u+1-\rho}\right) \frac{\rho(x+v\alpha)(1-\rho)}{(\rho u+1-\rho)^2} (1-u)^{\frac{1}{1-\rho}} \\
&\quad = \int_0^1 \mathrm{d}u\, (1-u)^{\frac{1}{1-\rho}} \int_0^{\frac{cx}{u}} \mathrm{d}v\, b\left(\frac{(x+v\alpha)(1-\rho)}{\rho u+1-\rho}\right) \frac{\rho(x+v\alpha)(1-\rho)}{(\rho u+1-\rho)^2} \\
&\quad = \frac{1}{\alpha(1-\rho)} \int_0^1 \mathrm{d}u\, (1-u)^{\frac{1}{1-\rho}} \int_{\frac{x(1-\rho)}{\rho u+1-\rho}}^{\frac{x(1-\rho)}{\rho u}} \mathrm{d}t\, t b(t).
\end{aligned}
$$

Finally, focus on the last term of Theorem 3.2. We use the changes of variables $z \mapsto w = z - v\alpha(1-\rho)$ and then $w \mapsto u = x(1-\rho)/w$.

$$
\begin{aligned}
&\int_{cx}^{\infty} \mathrm{d}v \int_{(v\alpha+x)(1-\rho)}^{v\alpha} \mathrm{d}z\, b(z) \left(1 - \frac{x(1-\rho)}{z - v\alpha(1-\rho)}\right)^{\frac{1}{1-\rho}} \\
&\quad = \int_{cx}^{\infty} \mathrm{d}v \int_{x(1-\rho)}^{\rho v\alpha} \mathrm{d}w\, b\big(w + v\alpha(1-\rho)\big) \left(1 - \frac{x(1-\rho)}{w}\right)^{\frac{1}{1-\rho}} \\
&\quad = \int_{x(1-\rho)}^{\infty} \mathrm{d}w \left(1 - \frac{x(1-\rho)}{w}\right)^{\frac{1}{1-\rho}} \int_{\frac{w}{\rho\alpha}}^{\infty} \mathrm{d}v\, b\big(w + va(1-\rho)\big) \\
&\quad = \frac{1}{\alpha(1-\rho)} \int_{x(1-\rho)}^{\infty} \mathrm{d}w \left(1 - \frac{x(1-\rho)}{w}\right)^{\frac{1}{1-\rho}} \mathsf{P}\left(B > \frac{w}{\rho}\right) \\
&\quad = \frac{x}{\alpha} \int_0^1 \frac{1}{u^2} \mathsf{P}\left(B > \frac{x(1-\rho)}{\rho u}\right) (1-u)^{\frac{1}{1-\rho}} du. \qquad (5.1)
\end{aligned}
$$

The proof of the lemma is completed by collecting the terms. □

Remark 5.1. It is interesting to note that (5.1) can be rewritten as follows:

$$\begin{aligned}
&\frac{x}{\alpha}\int_0^1 \frac{1}{u^2}\mathsf{P}\left(B > \frac{x(1-\rho)}{\rho u}\right)(1-u)^{\frac{1}{1-\rho}}du \\
&\quad = \frac{x}{\alpha}\int_0^1 \frac{1}{u^2}\mathsf{P}\left(B > \frac{x(1-\rho)}{\rho u}\right)\mathsf{P}\Big(W^*_{\rm ROS} > \frac{\beta u}{1-\rho}\Big)du \\
&\quad = \frac{x}{\alpha}\frac{\rho}{x(1-\rho)}\int_{\frac{x(1-\rho)}{\rho}}^{\infty}\mathsf{P}(B>z)\mathsf{P}\Big(W^*_{\rm ROS} > \frac{\alpha x}{z}\Big)dz \\
&\quad = \frac{\rho^2}{1-\rho}\mathsf{P}\Big(B^{fw} > \frac{x(1-\rho)}{\rho}, B^{fw}W^*_{\rm ROS} > \alpha x\Big).
\end{aligned}$$

In the case of Poisson arrivals and regularly varying service times, assuming again that

$$\mathsf{P}\left(B>x\right) \sim \frac{C}{\Gamma(1-\nu)}x^{-\nu}L(x),$$

and recalling that $\lambda = 1/\alpha$, it is easy to go from the expression in Lemma 5.1 to the expression

$$\mathsf{P}(W_{\rm ROS} > x) \quad \sim \quad \frac{CL(x)x^{1-\nu}}{\alpha(1-\rho)\Gamma(2-\nu)}\left[1-\rho+\int_0^1 g(u,\rho,\nu)\mathrm{d}u\right].$$

Using (4.16) it is seen that this corresponds to Theorem 4.1.

6 Heavy-traffic limit for the waiting time distribution

In this section, we consider the $M/G/1$ queue with general service time distribution. We are interested in the heavy-traffic case: $\rho \to 1$. The main result is to find a sequence $\Delta(\rho)$ which tends to 0 as $\rho \to 1$ such that $\mathsf{E}[\mathrm{e}^{-\omega\Delta(\rho)W_{\rm ROS}}]$ tends to a proper limit as $\rho \to 1$. This way we shall be able to retrieve a result of Kingman [20] for the case of finite service time variance, as well as derive a new result for the case of regularly varying service time distribution with infinite variance.

The following lemma will be useful in the sequel.

Lemma 6.1. *The following bound holds for all $s > 0$ and $0 \le z \le 1$:*

$$\Psi(s,z) \le \exp\Big[-\frac{1-z}{\beta s}\Big]. \tag{6.1}$$

Moreover, for any $t > 0$,

$$\lim_{(\rho,s)\to(1,0)} \Psi(s, 1-\beta s t) = \mathrm{e}^{-t}. \tag{6.2}$$

Proof. Using the inequality $\beta\{s\} \ge 1-\beta s$, one finds

$$y - \beta\{s+\lambda(1-y)\} \le y - 1 + \beta s + \rho(1-y) \le \beta s,$$

and (6.1) follows from the definition (4.3), since

$$-\log\Psi(s,z) \ge \int_z^1 \frac{dy}{\beta s} = \frac{1-z}{\beta s}.$$

The proof of (6.2) follows a similar argument: indeed,

$$-\log\Psi(s,1-\beta st)=\int_{1-\beta st}^{1}\frac{\mathrm{d}y}{y-\beta\{s+\lambda(1-y)\}}=\int_{0}^{1}\frac{\beta st\mathrm{d}v}{1-\beta stv-\beta\{s+\rho stv\}},$$

and, as $s\to 0$,

$$1-\beta stv-\beta\{s+\rho stv\}\sim\beta s(1-(1-\rho)tv).$$

This yields (6.2) and concludes the proof of the lemma. □

The LST of the steady-state waiting time distribution under FCFS is given by (cf. [13], p. 255): for any $s\geq 0$,

$$\mathsf{E}[\mathrm{e}^{-sW_{\mathrm{FCFS}}}]=\frac{1-\rho}{1-\rho\beta^{fw}\{s\}}.$$

The following lemma illustrates the relation between W_{ROS} and W_{FCFS} in heavy traffic.

Lemma 6.2. *Assume that $\beta<\infty$ and that there exists $\Delta(\rho)>0$ which can serve as a proper scaling for W_{FCFS}: for any $\omega>0$,*

$$\lim_{\rho\to 1}\mathsf{E}[\mathrm{e}^{-\omega\Delta(\rho)W_{\mathrm{FCFS}}}]=\mathsf{E}[\mathrm{e}^{-\omega\widehat{W}_{\mathrm{FCFS}}}],\tag{6.3}$$

where $\widehat{W}_{\mathrm{FCFS}}$ is a non-negative random variable. Then $\Delta(\rho)$ is a proper scaling for W_{ROS}: for any $\omega>0$,

$$\lim_{\rho\to 1}\mathsf{E}[\mathrm{e}^{-\omega\Delta(\rho)W_{\mathrm{ROS}}}]=\int_{0}^{\infty}\mathsf{E}[\mathrm{e}^{-\omega t\widehat{W}_{\mathrm{FCFS}}}]\mathrm{e}^{-t}\mathrm{d}t\stackrel{\mathrm{def}}{=}\mathsf{E}[\mathrm{e}^{-\omega\widehat{W}_{\mathrm{ROS}}}].$$

Proof. The starting point of the proof is Equation (4.6). First, using integration by parts,

$$\int_{\mu\{s\}}^{1}\frac{1-z-\frac{\beta s}{1-\rho}}{z-\beta\{s+\lambda(1-z)\}}\Psi(s,z)\mathrm{d}z=-\frac{\beta s}{1-\rho}+\int_{\mu\{s\}}^{1}\Psi(s,z)\mathrm{d}z,$$

and the integral above can be bounded using (6.1) as

$$\int_{\mu\{s\}}^{1}\Psi(s,z)\mathrm{d}z\leq\int_{\mu\{s\}}^{1}\exp\Big[-\frac{1-z}{\beta s}\Big]\mathrm{d}z\leq\beta s.$$

The second part of (4.6) can be expressed in terms of W_{FCFS}:

$$\begin{aligned}\int_{\mu\{s\}}^{1}&\frac{1-z}{z-\beta\{\lambda(1-z)\}}\Psi(s,z)\mathrm{d}z\\&=-\int_{\mu\{s\}}^{1}\frac{1}{1-\rho}\mathsf{E}[\mathrm{e}^{-\lambda(1-z)W_{\mathrm{FCFS}}}]\Psi(s,z)\mathrm{d}z\\&=-\frac{\beta s}{1-\rho}\int_{0}^{\frac{\varepsilon(s)}{\beta s}}\mathsf{E}[\mathrm{e}^{-\rho stW_{\mathrm{FCFS}}}]\Psi(s,1-\beta st)\mathrm{d}t.\end{aligned}$$

Plugging these two relations into (4.6) yields

$$\mathsf{E}[\mathrm{e}^{-sW_{\mathrm{ROS}}}]=\rho\int_{0}^{\frac{\varepsilon(s)}{\beta s}}\mathsf{E}[\mathrm{e}^{-\rho stW_{\mathrm{FCFS}}}]\Psi(s,1-\beta st)\mathrm{d}t+O(1-\rho),\tag{6.4}$$

uniformly in $s > 0$. We now show that

$$\lim_{\rho \to 1} \frac{\varepsilon(\omega \Delta(\rho))}{\Delta(\rho)} = +\infty, \tag{6.5}$$

with, as before, $\varepsilon(s) \stackrel{\text{def}}{=} 1 - \mu\{s\}$ and $\mu\{s\}$ is the LST of the busy period distribution. Let $\mu_{\bar{\rho}}(s)$ be the LST of the busy period of the $M/G/1$ with service time distributions $B(\cdot)$ and load $\bar{\rho}$. Obviously, for $\rho \geq \bar{\rho}$ we have $\mu(s) \leq \mu_{\bar{\rho}}(s)$, for all $s > 0$. Hence,

$$\liminf_{\rho \to 1} \frac{\varepsilon(\omega \Delta(\rho))}{\Delta(\rho)} \geq \lim_{\rho \to 1} \frac{1 - \mu_{\bar{\rho}}(\omega \Delta(\rho))}{\Delta(\rho)} = \frac{\beta \omega}{1 - \bar{\rho}}.$$

This is true for any fixed $\bar{\rho} \in (0, 1)$. Letting $\bar{\rho}$ pass to 1 we obtain (6.5).

Finally, under Assumption (6.3), Lemma 6.1 allows to apply the Dominated Convergence Theorem to (6.4), and the lemma is proved. □

Using Feller's continuity theorem, Lemma 6.2 can be rewritten in a more compelling way.

Corollary 6.3. *Assume that there exists $\Delta(\rho) > 0$, and a random variable $\widehat{W}_{\mathrm{FCFS}}$, such that the following limit holds in distribution:*

$$\lim_{\rho \to 1} \Delta(\rho) W_{\mathrm{FCFS}} = \widehat{W}_{\mathrm{FCFS}}.$$

Then, in distribution,

$$\lim_{\rho \to 1} \Delta(\rho) W_{\mathrm{ROS}} = Y \widehat{W}_{\mathrm{FCFS}} \stackrel{\text{def}}{=} \widehat{W}_{\mathrm{ROS}}, \tag{6.6}$$

where Y is an exponential random variable with mean 1, independent of $\widehat{W}_{\mathrm{FCFS}}$.

Remark 6.1. In view of the PASTA property and the fact that the workload is the same under FCFS and ROS, (6.6) states that the scaled waiting time $\widehat{W}_{\mathrm{ROS}}$ equals in distribution the product of the unit exponential Y and the scaled workload $\widehat{V}_{\mathrm{ROS}} \stackrel{\mathrm{d}}{=} \widehat{W}_{\mathrm{FCFS}}$. Put differently, $\mathsf{P}(\hat{W}_{\mathrm{ROS}} > x | \hat{V}_{\mathrm{ROS}} = y) = \mathsf{P}(Y > x/y) = \mathrm{e}^{-x/y}$. The latter result might be intuitively understood by referring to a *snapshot principle*. Consider a tagged customer. If it is not being elected for service at the end of some ROS services, then in the heavy-traffic limit the remaining workload that it sees has not changed. The randomness ('memoryless') property of ROS then implies that the remaining waiting time of the tagged customer has the same distribution as before.

As a first application of Lemma 6.2, consider the case where the service time distribution has a finite second moment. The following theorem extends a result of Kingman [20], where it was additionally assumed that $\beta\{s\}$ exists for some $s < 0$.

Theorem 6.4. *Assume that the variance σ^2 of the service time is finite and let*

$$\Delta(\rho) = \frac{\lambda(1 - \rho)}{1 + \frac{1}{2}\lambda^2 \sigma^2}. \tag{6.7}$$

Then, for any $\omega > 0$,

$$\lim_{\rho \to 1} \mathsf{E}[\mathrm{e}^{-\omega \Delta(\rho) W_{\mathrm{ROS}}}] = \int_0^\infty \frac{\mathrm{e}^{-t}}{1 + \omega t} \mathrm{d}t.$$

Proof. It is known from a classical result of Kingman [19] that, with $\Delta(\rho)$ defined by (6.7),

$$\lim_{\rho\to 1} \mathsf{E}[\mathrm{e}^{-\omega\Delta(\rho)W_{\mathrm{FCFS}}}] = \frac{1}{1+\omega},$$

so that we may apply Lemma 6.2. □

Remark 6.2. Thus, when the service times have a finite variance, $\widehat{W}_{\mathrm{FCFS}}$ has an exponential distribution, so that $\widehat{W}_{\mathrm{ROS}}$ is the product of two independent exponentials with unit mean. Letting $K_1(\cdot)$ be the modified Bessel function of the second kind, simple calculations show that

$$\begin{aligned}\mathsf{P}(\widehat{W}_{\mathrm{ROS}} > x) &= \int_0^\infty \mathsf{P}(t\widehat{W}_{\mathrm{FCFS}} > x)\mathrm{e}^{-t}\mathrm{d}t \\ &= \int_0^\infty \exp\Big[-\frac{x}{t} - t\Big]\mathrm{d}t \\ &= 2\sqrt{x}K_1(2\sqrt{x}),\end{aligned}$$

which coincides with Theorem 6 of [20].

In the case of a service time distribution with regularly varying tail and infinite variance a similar, but new, result can be obtained from results of Boxma and Cohen [8] for FCFS.

Theorem 6.5. *Under Assumption (4.8), with $1 < \nu < 2$, let $\Delta(\rho)$ be the unique root of the equation*

$$\lambda C x^{\nu-1} L(x) = 1 - \rho,\ x > 0, \tag{6.8}$$

such that $\Delta(\rho) \downarrow 0$ for $\rho \uparrow 1$. Then

$$\lim_{\rho\to 1} \mathsf{E}[\mathrm{e}^{-\omega\Delta(\rho)W_{\mathrm{ROS}}}] = \int_0^\infty \frac{\mathrm{e}^{-t}}{1+(\omega t)^{\nu-1}}\mathrm{d}t.$$

Proof. It has been proved in [8] that $\Delta(\rho)$ exists and that for any $\omega > 0$, under Assumption (4.8),

$$\lim_{\rho\to 1} \mathsf{E}[\mathrm{e}^{-\omega\Delta(\rho)W_{\mathrm{FCFS}}}] = \frac{1}{1+\omega^{\nu-1}}.$$

The theorem now follows from Lemma 6.2. □

Appendices

A Classes of distributions

Definitions and properties

We say that a random variable belongs to a certain class if its distribution function belongs to that class.

1. A cdf F belongs to the class $\mathcal{L}$ of long-tailed distributions if there exists a $y > 0$ (or, equivalently, for all $y > 0$) such that, as $x \to \infty$,

$$\frac{\overline{F}(x+y)}{\overline{F}(x)} \to 1.$$

(a) If $F \in \mathcal{L}$, $c > 0$, and G is another distribution such that $\overline{G}(x) \sim c\overline{F}(x)$ as $x \to \infty$, then $G \in \mathcal{L}$.

(b) If $F \in \mathcal{L}$ and $m^+ \equiv m^+(F) = \int_0^\infty \overline{F}(t)dt$ is finite, then the integrated tail distribution F^I belongs to $\mathcal{L}$ too, but the converse is not true, in general. Here

$$F^I(x) = \max\left(0; 1 - \int_x^\infty \overline{F}(t)dt\right).$$

(c) $F^I \in \mathcal{L}$ if and only if $\overline{F}(x) = o(\overline{F}^I(x))$ as $x \to \infty$.

2. A cdf F belongs to the class $\mathcal{RV}$ of regularly varying distributions if there exists a $\nu > 0$ such that

$$\overline{F}(x) = x^{-\nu}L(x),$$

where $L(x)$ is a slowly varying (at infinity) function.

3. A cdf F belongs to the class $\mathcal{D}$ if

$$\inf_{x \geq 0} \frac{\overline{F}(2x)}{\overline{F}(x)} > 0.$$

(a) If $F \in \mathcal{D}$ and $m^+ < \infty$, then $F^I \in \mathcal{D}$.

4. A cdf F belongs to the class $\mathcal{IRV}$ of intermediate regularly varying distributions if

$$\lim_{\zeta \downarrow 1} \liminf_{x \to \infty} \frac{\overline{F}(\zeta x)}{\overline{F}(x)} = 1.$$

5. A cdf F on the positive half-line belongs to the class $\mathcal{S}$ of subexponential distributions if

$$\int_0^x F(dt)\overline{F}(x-t) \sim \overline{F}(x) \quad \text{as} \quad x \to \infty.$$

(a) A cdf F on the real line belongs to $\mathcal{S}$ if $F(x)\mathbf{I}(x \geq 0)$ belongs to $\mathcal{S}$.

6. A cdf F belongs to the class $\mathcal{S}^*$ if m^+ is finite and

$$\int_0^x \overline{F}(y)\overline{F}(x-y)dy \sim 2m^+\overline{F}(x) \quad \text{as} \quad x \to \infty.$$

7. Relations

(a) [14, p. 50] $\mathcal{RV} \subset \mathcal{IRV} \subset \mathcal{L} \bigcap \mathcal{D} \subset \mathcal{S}$.

(b) [21] If $F \in \mathcal{S}^*$, then $F \in \mathcal{S}$ and $F^I \in \mathcal{S}$.

(c) [21] If $F \in \mathcal{L} \bigcap \mathcal{D}$ and if m^+ is finite, then $F \in \mathcal{S}^*$.

(d) If $F \in \mathcal{D}$ and F has an eventually non-increasing density, then $F \in \mathcal{IRV}$. Indeed, put

$$K = \sup_{x \geq 0} \frac{\overline{F}(x/2)}{\overline{F}(x)}.$$

Then, for $\zeta > 1$ and for all sufficiently large x,

$$\frac{\overline{F}(x) - \overline{F}(\zeta x)}{\overline{F}(x)} = \frac{\overline{F}(x/2)}{\overline{F}(x)} \cdot \frac{\overline{F}(x) - \overline{F}(\zeta x)}{\overline{F}(x/2)}$$
$$\leq K \frac{\int_x^{\zeta x} f(t)dt}{\int_{x/2}^{x} f(t)dt} \leq K \frac{(\zeta - 1)xf(x)}{\frac{1}{2}xf(x)} = 2K(\zeta - 1) \to 0$$

as $\zeta \downarrow 1$.

(e) It follows from Properties (3a) and (7d) that if $F \in \mathcal{D}$ and $m^+ < \infty$, then $F^I \in \mathcal{IRV}$.

There exists a very useful relation between the tail behavior of a regularly varying probability distribution and the behavior of its LST near the origin. That relation often enables one to conclude from the form of the LST of a distribution, that the distribution itself is regularly varying at infinity. We present this relation in Lemma A.1 below. We use this in Section 4 to prove that the waiting time distribution in the $M/G/1$ queue under the ROS discipline is regularly varying at infinity if the service time distribution is regularly varying at infinity.

Let $F(\cdot)$ be the distribution of a non-negative random variable, with LST $\phi\{s\}$ and finite first n moments $\mu_1, \ldots, \mu_n$ (and $\mu_0 = 1$). Define

$$\phi_n\{s\} \stackrel{\text{def}}{=} (-1)^{n+1}\Big[\phi\{s\} - \sum_{j=0}^{n} \mu_j \frac{(-s)^j}{j!}\Big].$$

Lemma A.1. *Let $n < \nu < n+1$, $C \geq 0$. The following statements are equivalent:*

$$\phi_n\{s\} = (C + o(1))s^\nu L(1/s), \quad s \downarrow 0, \quad s \text{ real},$$

$$1 - F(x) = (C + o(1))\frac{(-1)^n}{\Gamma(1-\nu)} x^{-\nu} L(x), \quad x \to \infty.$$

The case $C > 0$ is due to Bingham and Doney [4]. The case $C = 0$ was first obtained by Vincent Dumas, and is treated in [10], Lemma 2.2. The case of an integer ν is more complicated; see Theorem 8.1.6 and Chapter 3 of [5].

B Proof of Theorem 2.4

Note that the distribution of the residual service time of the customer in service is the same for all non-preemptive and non-idling service disciplines. We may therefore concentrate on the FCFS discipline.

As before, V_{-n} denotes the amount of work in the system upon arrival of customer $-n$ and T_{-n} is the time between arrival of customer $-n$ and time 0 (which is the arrival time of customer 0). In the sequel the random variable V has the stationary workload distribution.

Lower bound. For any $\varepsilon \in (0,1)$, choose $K_1 > 0$ such that $\mathsf{P}(V > K_1) \leq \varepsilon$. Then choose an integer $n > 0$ such that $\mathsf{P}(T_{-n} > K_1) \geq 1 - \varepsilon$. Third, choose $K_2 > K_1$ such that

$\mathsf{P}(T_{-n} \in (K_1, K_2)) \geq 1 - 2\varepsilon$. Then, as $x \to \infty$,

$$\begin{aligned}
\mathsf{P}(B^{rp} > x) &\geq \sum_{m=n}^{\infty} \mathsf{P}(V_{-m} \leq K_1, T_{-n} \in (K_1, K_2), B_{-m} > x + T_{-m}) \\
&\geq \sum_{m=n}^{\infty} \mathsf{P}(V_{-m} \leq K_1, T_{-n} \in (K_1, K_2), B_{-m} > x + T_{-m} - T_{-n} + K_2) \\
&\geq (1-2\varepsilon)(1-\varepsilon) \sum_{l=0}^{\infty} \mathsf{P}(B_0 > x + K_2 + T_{-l}) \\
&\geq (1-2\varepsilon)(1-\varepsilon)\rho \mathsf{P}(B^{fw} > x + K_2) \\
&\sim (1-2\varepsilon)(1-\varepsilon)\rho \mathsf{P}(B^{fw} > x).
\end{aligned}$$

Letting $\varepsilon \to 0$, we get the correct lower bound. Since $B \in \mathcal{D}$, the lower bound is of order $O(\mathsf{P}(Z^{fw} > cx))$ for any positive c.

Upper bound. Let y be a positive number and $\eta(y) = \min\{n \geq 1 \ : \ \sum_{i=1}^{n} B_i > y\}$, $\chi(y) = \sum_1^{\eta(y)} B_i - y$. Since EB is finite, it follows from basic renewal theory that the family of distributions of random variables $\{\chi(y),\ y > 0\}$ is tight, i.e. $u(x) \equiv \sup_{y>0} \mathsf{P}(\chi(y) > x) \to 0$ as $x \to \infty$.

Since $B^{rp} \leq Z^{rp}$ almost surely, we have by Corollary 2.3 and by the lower bound obtained,

$$\begin{aligned}
\mathsf{P}(B^{rp} > x) &= \mathsf{P}\left(B^{rp} > x, Z^{rp} > x\right) \\
&= (1 + o(1))\mathsf{P}(B^{rp} > x, Z^{rp} > x, \bigcup_{m=1}^{\infty} \{B_{-m} > (x + m\alpha)(1-\rho)\}) \\
&= (1 + o(1)) \sum_{m=1}^{\infty} \mathsf{P}(B^{rp} > x, Z^{rp} > x, B_{-m} > (x + m\alpha)(1-\rho)).
\end{aligned}$$

Denote by $f_m(x)$ the m-th term in the latter sum. For any $\varepsilon \in (0,1)$, choose $K > 0$ such that $\mathsf{P}(V > K) \leq \varepsilon$. Then

$$\begin{aligned}
f_m(x) &\leq \mathsf{P}(V_{-m} \leq K, B_{-m} > x + T_{-m} - K) \\
&+ \mathsf{P}(V_{-m} > K, B_{-m} > (x + m\alpha)(1-\rho)) \\
&+ \mathsf{P}(B^{rp} > x, V_{-m} \leq K, B_{-m} \in ((x + m\alpha)(1-\rho) - K, x + T_{-m} - K);\ \exists 1 \leq l \leq m \ : \ V_{-l} = 0) \\
&+ \mathsf{E} \int_{(x+m\alpha)(1-\rho)}^{x+T_{-m}-0} \mathsf{P}(V_{-m} \leq K, V_{-m} + B_{-m} \in dt) \mathsf{P}(\chi(x + T_{-m} - t) > x \mid T_{-m}) \\
&\equiv f_{m,1}(x) + f_{m,2}(x) + f_{m,3}(x) + f_{m,4}(x).
\end{aligned}$$

Here

$$\begin{aligned}
f_{m,1}(x) &\leq \mathsf{P}(B_{-m} > x + T_{-m} - K); \\
f_{m,2}(x) &\leq \varepsilon \mathsf{P}(B_{-m} > (x + m\alpha)(1-\rho)); \\
f_{m,3}(x) &\leq \sum_{l=1}^{m-1} \mathsf{P}(B_{-m} > (x + m\alpha)(1-\rho), V_{-l} = 0, V_{-j} > 0 \ \ \forall\, l < j < m) \mathsf{P}(B^{rp} > x \mid V_{-l} = 0) \\
&\leq \frac{\mathsf{P}(Z^{rp} > x)}{\mathsf{P}(V = 0)} \mathsf{P}(B_{-m} > (x + m\alpha)(1-\rho)); \\
f_{m,4}(x) &\leq u(x) \mathsf{P}(B_{-m} > (x + m\alpha)(1-\rho)).
\end{aligned}$$

Since $u(x) \to 0$ and $\mathsf{P}(Z^{rp} > x) \to 0$ as $x \to \infty$,

$$\sum_{m=1}^{\infty} (f_{m,2}(x) + f_{m,3}(x) + f_{m,4}(x)) \leq (1 + o(1))\varepsilon \mathsf{P}(Z^{rp} > x).$$

Further,

$$\sum_{m=1}^{\infty} f_{m,1}(x) \leq (1 + o(1))\rho \mathsf{P}(B^{fw} > x - K) \sim \rho \mathsf{P}(B^{fw} > x).$$

Letting $\varepsilon \downarrow 0$, the proof is completed. □

C Proof of Theorem 3.2

We focus on the waiting time W_{ROS} of customer 0 arriving at time 0. Our proof consists of three main parts, each corresponding to a typical scenario in which the large waiting time arises. The intuition behind these typical scenarios was discussed below Theorem 3.2 (it is convenient to treat the two "middle terms" as one scenario).

Before proceeding, we note that the distribution of the waiting time of customer 0 is not affected if we choose to use the FCFS discipline before time 0 and the ROS discipline after time 0. Thus, $W_{\mathrm{ROS}} \stackrel{\mathrm{d}}{=} W'_{\mathrm{ROS}}$, where W'_{ROS} denotes the waiting time of customer 0 under the modified service discipline.

The starting point of the proof is (2.14) in Corollary 2.5, which we repeat for convenience (the modified service discipline is non-idling and non-preemptive):

$$\mathsf{P}(W'_{\mathrm{ROS}} > x) \sim \sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (x + m\alpha)(1 - \rho)\right), \qquad x \to \infty.$$

In the verbal discussion, we interpret this relation as follows: there is one "large customer", i.e., customer $-m$ for which $B_{-m} > (x + m\alpha)(1 - \rho)$, that causes the large waiting time of customer 0. Note that any scenario in which the service of this large customer did not start before time 0 may be neglected (i.e., is of the order $o(\mathsf{P}(B^{fw} > x))$):

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (x + m\alpha)(1 - \rho), T_{-m} \leq V_{-m}\right) \\
&\leq \sum_{m=1}^{\infty} \mathsf{P}\left(B_{-m} > (x + m\alpha)(1 - \rho), T_{-m} \leq V_{-m}\right) \\
&\leq \sum_{m=1}^{\infty} \mathsf{P}\left(B_{-m} > (x + m\alpha)(1 - \rho), T_{-m} \leq V_{-m}, V_{-m} > K\right) \\
&\quad + \sum_{m=1}^{\infty} \mathsf{P}\left(B_{-m} > (x + m\alpha)(1 - \rho), T_{-m} \leq V_{-m} \leq K\right) \\
&\leq \mathsf{P}(V > K) \sum_{m=1}^{\infty} \mathsf{P}\left(B_{-m} > (x + m\alpha)(1 - \rho)\right) \\
&\quad + \mathsf{P}(T_{-M} \leq K) \sum_{m>M}^{\infty} \mathsf{P}\left(B_{-m} > (x + m\alpha)(1 - \rho)\right) + o(\mathsf{P}(B^{fw} > x)) \\
&= \left(\mathsf{P}(V > K) + \mathsf{P}(T_{-M} \leq K)\right) \frac{\rho}{1 - \rho} \mathsf{P}(B^{fw} > x),
\end{aligned}$$

which may be neglected after first taking $M \to \infty$ and then $K \to \infty$. We have used that the sum of a finite number of terms in the above summations is of the order $o(\mathsf{P}(B^{fw} > x))$, see Property (1c) in Appendix A. This property, as well as other steps taken in the proof of Theorem 2.4, will be used frequently in the following. We have thus proved that, as $x \to \infty$,

$$\mathsf{P}(W'_{\rm ROS} > x) \sim \sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\rm ROS} > x, B_{-m} > (x+m\alpha)(1-\rho), V_{-m} < T_{-m}\right) + o(\mathsf{P}(B^{fw} > x)).$$

Part I. We start with the scenario that the large customer is in service for the entire interval $(0, x)$. This is the case when the workload $V_{-m} < T_{-m}$ and $B_{-m} > T_{-m} - V_{-m} + x$. This immediately implies that the waiting time of customer 0 exceeds x. By Theorem 2.4, we have:

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\rm ROS} > x, B_{-m} > (x+m\alpha)(1-\rho), V_{-m} < T_{-m}, B_{-m} > T_{-m} - V_{-m} + x\right) \\
&= \sum_{m=1}^{\infty} \mathsf{P}\left(B_{-m} > (x+m\alpha)(1-\rho), V_{-m} < T_{-m}, B_{-m} > T_{-m} - V_{-m} + x\right) \\
&\sim \mathsf{P}(B^{rp} > x).
\end{aligned}$$

This corresponds to the first term in Theorem 3.2.

Part II. We now investigate the event that $W'_{\rm ROS} > x$ occurs while the large customer (customer $-m$) is still in service at time 0, but not anymore at time x. Thus, $V_{-m} < T_{-m} < B_{-m} + V_{-m} < T_{-m} + x$. Let $W'_{\rm ROS}(q)$ be the remaining waiting time of customer 0 after the first service completion after time 0 if q is the number of competing customers at that instant. In the sequel we shall simply write $W'_{\rm ROS}(q)$ instead of $W'_{\rm ROS}(\lfloor q \rfloor)$ when q is not an integer. If customer $-m$ is still in service at time 0, we may write

$$W'_{\rm ROS} = V_{-m} + B_{-m} - T_{-m} + W'_{\rm ROS}(A(T_{-m}, T_{-m} + V_{-m} + B_{-m})),$$

where $A(s,t)$ denotes the number of arrivals between times s and t. We obviously have

$$\begin{aligned}
W'_{\rm ROS} &\geq B_{-m} - T_{-m} + W'_{\rm ROS}(A(T_{-m}, T_{-m} + B_{-m})), \\
W'_{\rm ROS} &\leq B_{-m} + W'_{\rm ROS}(A(T_{-m}, T_{-m} + V_{-m} + B_{-m})).
\end{aligned} \tag{C.1}$$

The first bound gives

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\rm ROS} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \leq T_{-m} + x\right) \\
&\geq \sum_{m=1}^{\infty} \mathsf{P}\left(B_{-m} - T_{-m} + W'_{\rm ROS}(A(T_{-m}, T_{-m} + B_{-m})) > x, B_{-m} > (m\alpha + x)(1-\rho)\ ,\right. \\
&\qquad \left. V_{-m} < T_{-m} < B_{-m} + V_{-m} \leq T_{-m} + x\right) \\
&\geq (1-\delta)^2 \sum_{m=M}^{\infty} \mathsf{P}\big(B_{-m} - m\alpha(1+\varepsilon) + W'_{\rm ROS}(\frac{1}{\alpha}(1-\varepsilon)B_{-m}) > x, B_{-m} > (m\alpha + x)(1-\rho), \\
&\qquad\qquad m\alpha(1+\varepsilon) < B_{-m} \leq m\alpha(1-\varepsilon) - K + x\big),
\end{aligned} \tag{C.2}$$

where, for fixed $\varepsilon > 0$, $\delta > 0$, we have chosen $K > 0$ such that $\mathsf{P}(V > K) < \delta$ and $M > 0$ such that $M\alpha(1-\varepsilon) > K$ and, for all $m \geq M$ and $y \geq M\alpha(1-\rho)$,

$$\mathsf{P}\left((1-\varepsilon)m\alpha < T_{-m} < (1+\varepsilon)m\alpha, A(T_{-m}, T_{-m}+y) \geq \frac{1}{\alpha}y(1-\varepsilon)\right) \geq 1-\delta,$$

which is possible by the strong law of large numbers.

Note that the summation in (C.2) is actually truncated at $m = (x-K)/(2\varepsilon)$. For notation it is convenient to make the summation run from $m = 0$ to ∞. Adding the terms for $m < M$ in (C.2) causes an error of the order $o(\mathsf{P}(B^{fw} > x))$ and since

$$\begin{aligned}\sum_{m=0}^{\infty} \mathsf{P}(m\alpha(1-\varepsilon) + x - K < B_{-m} < m\alpha + x) &\leq \varepsilon O(\mathsf{P}(B^{fw} > x)),\\ \sum_{m=0}^{\infty} \mathsf{P}(B_{-m} > (m\alpha+x)(1-\rho), m\alpha < B_{-m} < m\alpha(1+\varepsilon)) &\leq \varepsilon O(\mathsf{P}(B^{fw} > x)),\end{aligned}$$

we have,

$$\begin{aligned}&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha+x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \leq T_{-m} + x\right)\\ &+\frac{\varepsilon}{(1-\delta)^2} O(\mathsf{P}(B^{fw} > x))\\ &\geq \sum_{m=0}^{\infty} \mathsf{P}\big(B_{-m} - m\alpha(1+\varepsilon) + W'_{\mathrm{ROS}}(\frac{1}{\alpha}(1-\varepsilon)B_{-m}) > x, B_{-m} > (m\alpha+x)(1-\rho),\\ &\qquad\qquad m\alpha < B_{-m} \leq m\alpha + x\Big).\end{aligned}$$

Furthermore, by Lemma 3.1 we can find x_0 such that for all $x > x_0$:

$$\begin{aligned}&\int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \mathrm{d}\mathsf{P}(B \leq z)\, \mathsf{P}\left(W'_{\mathrm{ROS}}(\frac{1}{\alpha}(1-\varepsilon)z) > x - z + m\alpha(1+\varepsilon)\right)\\ &\geq (1-\delta)\int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \mathrm{d}\mathsf{P}(B \leq z)\, \left(\left(1 - \frac{(1-\rho)(x-z+m\alpha(1+\varepsilon))}{\rho(1-\varepsilon)z}\right)^+\right)^{\frac{1}{1-\rho}}\\ &\geq (1-\delta)(1-\gamma)\int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \mathrm{d}\mathsf{P}(B \leq z)\, \left(\left(1 - \frac{(1-\rho)(x-z+m\alpha)}{\rho z}\right)^+\right)^{\frac{1}{1-\rho}},\end{aligned}$$

where $\gamma > 0$ depends on ε. In the last step we used that, as $\varepsilon \to 0$,

$$\frac{(1-\rho)(x-z+m\alpha(1+\varepsilon))}{\rho(1-\varepsilon)z} \to \frac{(1-\rho)(x-z+m\alpha)}{\rho z},$$

uniformly in z within the area of integration. (This can be seen, using that $z \geq (m\alpha+x)(1-\rho)$

and $z \geq m\alpha$.) So that, with $\delta \to 0$, $\varepsilon \to 0$ and then $\gamma \to 0$ we have:

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \leq T_{-m} + x\right) \\
&\geq \sum_{m=0}^{\infty} \int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \left(1 - \frac{(1-\rho)(x+m\alpha-z)}{\rho z}\right)^{\frac{1}{1-\rho}} d\mathsf{P}\left(B \leq z\right) \\
&\qquad + o(\mathsf{P}(B^{fw} > x)).
\end{aligned}$$

Next we derive an upper bound,

$$\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \leq T_{-m} + x\right) \\
&= \sum_{m=1}^{\infty} \mathsf{P}\left(V_{-m} + B_{-m} - T_{-m} + W'_{\mathrm{ROS}}(A(T_{-m}, T_{-m} + V_{-m} + B_{-m})) > x,\right. \\
&\qquad\qquad \left. B_{-m} > (m\alpha + x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \leq T_{-m} + x\right) \\
&\leq \mathsf{P}(V > K)\, O(\mathsf{P}(B^{fw} > x)) \\
&\quad + \sum_{m=1}^{\infty} \mathsf{P}\left(K + B_{-m} - T_{-m} + W'_{\mathrm{ROS}}(A(T_{-m}, T_{-m} + K + B_{-m})) > x,\right. \\
&\qquad\qquad \left. B_{-m} > (m\alpha + x)(1-\rho), T_{-m} - K < B_{-m} \leq T_{-m} + x\right) \\
&\leq \left(\delta + \mathsf{P}(V > K)\right)\, O(\mathsf{P}(B^{fw} > x)) \\
&\quad + \sum_{m=M}^{\infty} \mathsf{P}\Big(K + B_{-m} - m\alpha(1-\varepsilon) + W'_{\mathrm{ROS}}(\frac{1+\varepsilon}{\alpha}(K + B_{-m})) > x,\right. \\
&\qquad\qquad \left. B_{-m} > (m\alpha + x)(1-\rho), m\alpha(1-\varepsilon) - K < B_{-m} \leq m\alpha(1+\varepsilon) + x\Big),
\end{aligned}$$

where, for fixed $\varepsilon > 0$, we have chosen $M > 0$ such that, for all $m \geq M$ and $y \geq M\alpha(1-\rho)$,

$$\mathsf{P}\left((1-\varepsilon)m\alpha < T_{-m} < (1+\varepsilon)m\alpha, A(T_{-m}, T_{-m} + y) \leq \frac{1}{\alpha} y(1+\varepsilon)\right) \geq 1 - \delta.$$

As in the lower bound we may let the summation run from $m = 1$ to ∞ and replace the condition $m\alpha(1-\varepsilon) - K < B_{-m} \leq m\alpha(1+\varepsilon) + x$ with $m\alpha - K < B_{-m} \leq m\alpha + x - K$; the error we make is of the order $\varepsilon\, O(\mathsf{P}(B^{fw} > x))$. Also, replacing $B > (x + m\alpha)(1-\rho)$ by

$B > (x + m\alpha)(1-\rho) - K$ does not decrease the probability.

$$
\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \le T_{-m} + x\right) \\
&\le (\varepsilon + \delta + \mathsf{P}(V > K))\, O(\mathsf{P}(B^{fw} > x)) \\
&\quad + \sum_{m=0}^{\infty} \mathsf{P}\Big(K - m\alpha(1-\varepsilon) + B_{-m} + W'_{\mathrm{ROS}}(\frac{1+\varepsilon}{\alpha}(K + B_{-m})) > x, \\
&\qquad\qquad B_{-m} > (m\alpha + x)(1-\rho) - K, m\alpha - K < B_{-m} \le m\alpha + x - K\Big) \\
&= (\varepsilon + \delta + \mathsf{P}(V > K))\, O(\mathsf{P}(B^{fw} > x)) \\
&\quad + \sum_{m=0}^{\infty} \int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \mathrm{d}\mathsf{P}(B \le z - K)\, \mathsf{P}\Big(W'_{\mathrm{ROS}}(\frac{1+\varepsilon}{\alpha} z) > x + m\alpha(1-\varepsilon) - z\Big) \\
&\le (\varepsilon + \delta + \mathsf{P}(V > K))\, O(\mathsf{P}(B^{fw} > x)) \\
&\quad + (1+\gamma) \sum_{m=0}^{\infty} \int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \mathrm{d}\mathsf{P}(B \le z - K) \left(\left(1 - \frac{(1-\rho)(x + m\alpha - z)}{\rho z}\right)^{+}\right)^{\frac{1}{1-\rho}}.
\end{aligned}
$$

In the last step we use Lemma 3.1 and the uniform convergence of

$$\frac{(1-\rho)(x + m\alpha(1-\varepsilon) - z)}{\rho(1+\varepsilon)z}$$

as $\varepsilon \to 0$ ($\gamma > 0$ depends on ε).
Using that

$$\sum_{m=0}^{\infty} \mathsf{P}(\max\{(m\alpha + x)(1-\rho), m\alpha\} - K < B < \max\{(m\alpha + x)(1-\rho), m\alpha\}) = o(\mathsf{P}(B^{fw} > x)),$$

and

$$\sum_{m=0}^{\infty} \mathsf{P}(m\alpha + x - K < B < m\alpha + x) = o(\mathsf{P}(B^{fw} > x)),$$

we may replace $\mathrm{d}\mathsf{P}(B > z - K)$ by $\mathrm{d}\mathsf{P}(B > z)$. Now let $K \to \infty$, $\varepsilon \to 0$, $\delta \to 0$ and then $\gamma \to 0$ to conclude that

$$
\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \le T_{-m} + x\right) \\
&\le o(\mathsf{P}(B^{fw} > x)) + \sum_{m=0}^{\infty} \int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \mathrm{d}\mathsf{P}(B \le z) \left(1 - \frac{(1-\rho)(x + m\alpha - z)}{\rho z}\right)^{\frac{1}{1-\rho}}.
\end{aligned}
$$

Together with the lower bound, this shows

$$
\begin{aligned}
&\sum_{m=1}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} < T_{-m} < B_{-m} + V_{-m} \le T_{-m} + x\right) \\
&= o(\mathsf{P}(B^{fw} > x)) + \sum_{m=0}^{\infty} \int_{z=\max\{(m\alpha+x)(1-\rho), m\alpha\}}^{m\alpha+x} \mathrm{d}\mathsf{P}(B \le z) \left(1 - \frac{(1-\rho)(x + m\alpha - z)}{\rho z}\right)^{\frac{1}{1-\rho}}.
\end{aligned}
$$

The second and third term in Theorem 3.2 now readily follow, using that

$$\begin{aligned}
&\sum_{m=0}^{\infty} \int_{z=\max\{(m\alpha+x)(1-\rho),m\alpha\}}^{m\alpha+x} \mathrm{dP}(B \le z) \left(1 - \frac{(1-\rho)(x+m\alpha-z)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&= \int_{v=0}^{\infty} \mathrm{d}v \int_{z=\max\{(v\alpha+x)(1-\rho),v\alpha\}}^{v\alpha+x} \mathrm{dP}(B \le z) \left(1 - \frac{(1-\rho)(x+v\alpha-z)}{\rho z}\right)^{\frac{1}{1-\rho}} \\
&\quad + o(\mathsf{P}(B^{fw} > x)).
\end{aligned}$$

Part III. Finally, we deal with the last possible scenario in which service of the large customer ends before time 0. Thus, $V_{-m} + B_{-m} < T_{-m}$. Suppose that customer $-m+N$ is in service at time 0; $1 \le N \le m-1$. We can bound the waiting time of customer 0 from below by

$$W'_{\mathrm{ROS}} \;\ge\; W'_{\mathrm{ROS}}(A(-T_{-m},0) - N) \;=\; W'_{\mathrm{ROS}}(m-N),$$

and from above by

$$W'_{\mathrm{ROS}} \;\le\; B_{-m+N} + W'_{\mathrm{ROS}}(m-N+1).$$

We start with the lower bound. We shall denote the number of departures in the interval $[u,v)$ by $D(u,v)$. Note that $N = D(-T_{-m}+V_{-m}+B_{-m},0) \le D(-T_{-m}+B_{-m},0)$. In the following we take ε, δ, M and K such that $\mathsf{P}(V > K) < \delta$ and for all $m \ge M$, $y \ge K$,

$$\mathsf{P}\left((1-\varepsilon)m\alpha < T_{-m} < (1+\varepsilon)m\alpha, D(-y,0) > (1-\varepsilon)\frac{y}{\beta}\right) > 1-\delta.$$

We have,

$$\begin{aligned}
&\sum_{m=0}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha+x)(1-\rho), V_{-m}+B_{-m} < T_{-m}\right) \\
&\ge (1-\delta)^2 \sum_{m=M}^{\infty} \mathsf{P}\Big(W'_{\mathrm{ROS}}(m-(1-\varepsilon)\frac{1}{\beta}((1+\varepsilon)m\alpha - B_{-m})) > x, \\
&\qquad\qquad B_{-m} > (m\alpha+x)(1-\rho), K + B_{-m} < (1-\varepsilon)m\alpha\Big) \\
&= \varepsilon O(\mathsf{P}(B^{fw} > x)) + (1-\delta)^2 \sum_{m=0}^{\infty} \mathsf{P}\Big(W'_{\mathrm{ROS}}(m-(1-\varepsilon)\frac{1}{\beta}((1+\varepsilon)m\alpha - B_{-m})) > x, \\
&\qquad\qquad B_{-m} > (m\alpha+x)(1-\rho), B_{-m} < m\alpha\Big) \\
&= \varepsilon O(\mathsf{P}(B^{fw} > x)) \\
&\quad + (1-\delta)^2 \sum_{m=0}^{\infty} \int_{z=(m\alpha+x)(1-\rho)}^{m\alpha} \mathrm{dP}(B \le z)\, \mathsf{P}\Big(W'_{\mathrm{ROS}}(m-(1-\varepsilon)\frac{1}{\beta}((1+\varepsilon)m\alpha - z)) > x\Big) \\
&\ge \varepsilon O(\mathsf{P}(B^{fw} > x)) \\
&\quad + (1-\gamma)(1-\delta)^2 \sum_{m=0}^{\infty} \int_{z=(m\alpha+x)(1-\rho)}^{m\alpha} \mathrm{dP}(B \le z) \left(1 - \frac{(1-\rho)x}{z - m\alpha(1-\rho)}\right)^{\frac{1}{1-\rho}},
\end{aligned}$$

where $\gamma > 0$ depends on ε. In the last step we used Lemma 3.1 and the uniform convergence of

$$\frac{(1-\rho)x}{\beta(m-(1-\varepsilon)\frac{1}{\beta}\left((1+\varepsilon)m\alpha - z\right))}$$

as $\varepsilon \to 0$. Similar to Part II it can be shown that

$$\begin{aligned}
&\sum_{m=0}^{\infty} \int_{z=(m\alpha+x)(1-\rho)}^{m\alpha} \mathrm{dP}(B \le z) \left(1 - \frac{(1-\rho)x}{z - m\alpha(1-\rho)}\right)^{\frac{1}{1-\rho}} \\
&= o(\mathsf{P}(B^{fw} > x)) + \int_{v=0}^{\infty} \int_{z=(v\alpha+x)(1-\rho)}^{v\alpha} \mathrm{dP}(B \le z) \left(1 - \frac{(1-\rho)x}{z - v\alpha(1-\rho)}\right)^{\frac{1}{1-\rho}}.
\end{aligned}$$

Letting $\varepsilon \to 0$, $\delta \to 0$ and $\gamma \to 0$ we thus have proved that

$$\begin{aligned}
&\sum_{m=0}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} + B_{-m} < T_{-m}\right) \\
&\ge o(\mathsf{P}(B^{fw} > x)) + \int_{v=0}^{\infty} \int_{z=(v\alpha+x)(1-\rho)}^{v\alpha} \mathrm{dP}(B \le z) \left(1 - \frac{(1-\rho)x}{z - v\alpha(1-\rho)}\right)^{\frac{1}{1-\rho}}.
\end{aligned}$$

It remains to show that the right-hand side is also an upper bound for the left-hand side. Recall that $N = D(-T_{-m} + V_{-m} + B_{-m}, 0)$ and $W'_{\mathrm{ROS}} \le B_{-m+N} + W'_{\mathrm{ROS}}(m - N + 1)$. Note that, if $\varepsilon > 0$ and $\delta > 0$ and M such that $\mathsf{P}(T_{-m} > (1+\varepsilon)m\alpha) < \delta$ for all $m \ge M$, then

$$\begin{aligned}
&\sum_{m=0}^{\infty} \mathsf{P}\left(B_{-m} > (m\alpha + x)(1-\rho), V_{-m} + B_{-m} < T_{-m}, V_{-m} + B_{-m} > (1-2\varepsilon)m\alpha\right) \\
&\le (\delta + \mathsf{P}(V > K)) O(\mathsf{P}(B^{fw} > x)) \\
&\quad + \sum_{m=M}^{\infty} \mathsf{P}\left(B_{-m} > (m\alpha + x)(1-\rho), (1-2\varepsilon)m\alpha - K < B_{-m} < (1+\varepsilon)m\alpha\right) \\
&= (\varepsilon + \delta + \mathsf{P}(V > K)) O(\mathsf{P}(B^{fw} > x)).
\end{aligned}$$

We shall use this in what follows. In addition, let M and K be such that $\mathsf{P}(V > K) < \delta$, and for all $m \ge M$ and $y \ge \varepsilon M\alpha$,

$$\mathsf{P}(D(-y, 0) \le (1-\varepsilon)y/\beta) < \delta.$$

Also, we take L such that $\mathsf{P}(B > L) < \delta$.

$$
\begin{aligned}
&\sum_{m=0}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}} > x, B_{-m} > (m\alpha + x)(1-\rho), V_{-m} + B_{-m} < T_{-m}\right) \\
&\leq (\varepsilon + \delta + \mathsf{P}(V > K) + \mathsf{P}(B > L))O(\mathsf{P}(B^{fw} > x)) \\
&+ \sum_{m=M}^{\infty} \mathsf{P}\left(W'_{\mathrm{ROS}}(m - N + 1) > x - L, B_{-m} > (m\alpha + x - L)(1-\rho), V_{-m} + B_{-m} < (1-2\varepsilon)m\alpha\right) \\
&\leq (\varepsilon + \delta)O(\mathsf{P}(B^{fw} > x + L)) \\
&\quad + \sum_{m=M}^{\infty} \mathsf{P}\Big(W'_{\mathrm{ROS}}(m - (1-\varepsilon)\frac{1}{\beta}((1-\varepsilon)m\alpha - K - B_{-m}) + 1) > x, \\
&\qquad\qquad B_{-m} > (m\alpha + x)(1-\rho), B_{-m} < (1-2\varepsilon)m\alpha\Big) \\
&= (\varepsilon + \delta)O(\mathsf{P}(B^{fw} > x)) \\
&\quad + \sum_{m=0}^{\infty} \mathsf{P}\Big(W'_{\mathrm{ROS}}(m - (1-\varepsilon)\frac{1}{\beta}((1-\varepsilon)m\alpha - K - B_{-m}) + 1) > x, \\
&\qquad\qquad B_{-m} > (m\alpha + x)(1-\rho) - K - 1, B_{-m} < m\alpha - K - 1\Big) \\
&= (\varepsilon + \delta)O(\mathsf{P}(B^{fw} > x)) \; + \; (1+\gamma)\int_{v=0}^{\infty}\int_{z=(v\alpha+x)(1-\rho)}^{v\alpha} \mathrm{d}\mathsf{P}(B \leq z) \left(1 - \frac{(1-\rho)x}{z - v\alpha(1-\rho)}\right)^{\frac{1}{1-\rho}}.
\end{aligned}
$$

As before, in the last step we use Lemma 3.1, the uniform convergence of

$$\frac{(1-\rho)x}{\beta m - (1-\varepsilon)((1-\varepsilon)m\alpha - K - B_{-m}) + \beta},$$

as $\varepsilon \to 0$ and the fact that replacing the summation with an integral and $\mathrm{d}\mathsf{P}(B \leq z - K - 1)$ with $\mathrm{d}P(B \leq z)$ introduces an error of the order $o(\mathsf{P}(B^{fw} > x))$. Letting $\varepsilon \to 0$, $\delta \to 0$ and $\gamma \to 0$ yields the last term in Theorem 3.2. □

D Random and deterministic arrivals

The following lemma states that when interested in events involving a large service time, we may in fact ignore the randomness in the arrival process and replace it by a deterministic arrival process with the same mean arrival rate. Thus, heuristically, we may concentrate on the $D/G/1$ queue instead of the $GI/G/1$ queue. Although the lemma is not explicitly used in the paper, it has been very useful in guiding us to the proof of several of its results. We formulate it here because we expect that a reduction to a deterministic arrival process will often be helpful in proving tail asymptotics, see also Baccelli and Foss [2].

Lemma D.1. *If $B = B_0$ has a finite first moment and if its integrated tail distribution belongs*

to $\mathcal{L}$, then, for any constant $c > 0$, as $x \to \infty$,

$$\begin{aligned}\sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cT_{-m}\right) &\sim \sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cT_{-m}, B > x + cm\alpha\right) \\ &\sim \sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cm\alpha\right) \\ &\sim \frac{\rho}{c}\mathsf{P}\left(B^{fw} > x\right).\end{aligned}$$

Proof. From Appendix A (Property (1c) in Appendix A), the integrated tail distribution of B belongs to $\mathcal{L}$ if and only if $B^{fw} \in \mathcal{L}$.

Lower bound. For any $\varepsilon \in (0,1)$, choose $R > 0$ such that

$$\inf_{m \geq 0} \mathsf{P}\left(T_{-m} \leq m\alpha(1+\varepsilon) + R\right) \geq 1 - \varepsilon.$$

Then, as $x \to \infty$,

$$\begin{aligned}\sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cT_{-m}\right) &\geq \sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cm\alpha(1+\varepsilon) + cR, T_{-m} \leq m\alpha(1+\varepsilon) + R\right) \\ &\geq (1-\varepsilon)\sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cm\alpha(1+\varepsilon) + cR\right) \\ &\geq \frac{1-\varepsilon}{1+\varepsilon}\frac{\rho}{c}\mathsf{P}\left(B^{fw} > x + cR\right) \\ &\sim \frac{1-\varepsilon}{1+\varepsilon}\frac{\rho}{c}\mathsf{P}\left(B^{fw} > x\right).\end{aligned}$$

Letting $\varepsilon \downarrow 0$, we get the right lower bound.

Upper bound. Fix any $\varepsilon \in (0,1)$. Since T_{-m} are partial sums of non-negative i.i.d. r.v.'s with a finite positive mean α,

$$K \equiv \sum_{m=0}^{\infty} \mathsf{P}\left(T_{-m} \leq m\alpha(1-\varepsilon)\right) < \infty.$$

Then, as $x \to \infty$,

$$\begin{aligned}\sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cT_{-m}\right) &\leq \sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cm\alpha(1-\varepsilon), T_{-m} \geq m\alpha(1-\varepsilon)\right) \\ &+ \sum_{m=0}^{\infty} \mathsf{P}\left(B > x, T_{-m} \leq m\alpha(1-\varepsilon)\right) \\ &\leq \sum_{m=0}^{\infty} \mathsf{P}\left(B > x + cm\alpha(1-\varepsilon)\right) + K\mathsf{P}(B > x) \\ &\leq \frac{1}{1-\varepsilon}\frac{\rho}{c}\mathsf{P}\left(B^{fw} > x - c\alpha\right) + K\mathsf{P}(B > x) \\ &\sim \frac{1}{1-\varepsilon}\frac{\rho}{c}\mathsf{P}\left(B^{fw} > x\right),\end{aligned}$$

since $B^{fw} \in \mathcal{L}$ and $\mathsf{P}(B > x) = o(\mathsf{P}(B^{fw} > x))$, see Appendix A (Properties (1b) and (1c)). Letting $\varepsilon \downarrow 0$, we get an upper bound which coincides with the lower bound.

□

References

[1] Abramowitz, M., Stegun, I.A. (eds.) *Handbook of Mathematical Functions*. Dover Publications, Inc., New York, 1965.

[2] Baccelli, F., Foss, S.G. Moments and tails in monotone-separable stochastic networks. INRIA-ENS Report, 2001; to appear in Annals of Applied Probability.

[3] Bertsekas, D.P., Gallager, R.G. *Data Networks*. Prentice Hall, Englewood Cliffs (NJ), 1992.

[4] Bingham, N.H., Doney, R.A. Asymptotic properties of super-critical branching processes. I: The Galton-Watson process. *Adv. Appl. Probab. 6* (1974), 711–731.

[5] Bingham, N.H., Goldie, C.M., Teugels, J.L. *Regular Variation*. Cambridge University Press, Cambridge, UK, 1987.

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

[7] Borst, S.C., Boxma, O.J., Núñez Queija, R. Heavy tails: The effect of the service discipline. In *Computer Performance Evaluation*, Tony Field et al., eds. LNCS 2324, Springer, Berlin, 2002, pp. 1-30.

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

[9] Boxma, O.J., Denteneer, D., Resing, J.A.C. Some models for contention resolution in cable networks. In *Networking 2002*, E. Gregori et al., eds. LNCS 2345, Springer, Berlin, 2002, pp. 117-128.

[10] Boxma, O.J., Dumas, V. The busy period in the fluid queue. *Perf. Eval. Review 26* (1998), 100–110.

[11] Burke, P. Equilibrium delay distribution for one channel with constant holding time, Poisson input and random service. *Bell System Tech. J. 38* (1959), 1021-1031.

[12] Cohen, J.W. Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Probab. 10* (1973), 343–353.

[13] Cohen, J.W. *The Single Server Queue*. North-Holland, Amsterdam, 1982.

[14] Embrechts, P., Klüppelberg, C., Mikosch, T. *Modelling Extremal Events for Insurance and Finance*. Springer, Heidelberg, 1997.

[15] Flatto, L. The waiting time distribution for the random order service $M/M/1$ queue. *Ann. Appl. Probab. 7* (1997), 382-409.

[16] Foss, S., Zachary, S. Tail asymptotics for the busy cycle and the busy period in a single server queue with subexponential service time distribution. Working paper.

[17] Fuhrmann, S.W., Iliadis, I. A comparison of three random disciplines. *Queueing Systems 18* (1994), 249-271.

[18] Jelenković, P.R., Momcilović, P. Large deviations of square root insensitive random sums. Technical report 2002-05-101, Dept. of Electrical Engineering, Columbia University, 2002.

[19] Kingman, J.F.C. On queues in heavy traffic. *J. Roy. Statist. Soc. Ser. B 24* (1962), 383–392.

[20] Kingman, J.F.C. On queues in which customers are served in random order. *Proc. Cambridge Philos. Soc. 58* (1962), 79–91.

[21] Klüppelberg, C. Subexponential distributions and integrated tails. *J. Appl. Probab. 25* (1988), 132-141.

[22] Le Gall, P. *Les Systèmes avec ou sans Attente et les Processus Stochastiques*. Dunod, Paris, 1962.

[23] De Meyer, A., Teugels, J.L. On the asymptotic behaviour of the distributions of the busy period and service time in $M/G/1$. *J. Appl. Probab. 17* (1980), 802–813.

[24] Pakes, A.G. On the tails of waiting-time distributions. *J. Appl. Probab. 12* (1975), 555–564.

[25] Palm, C. Waiting times with random served queue. *Tele 1* (1957), 1-107 (English ed.; original from 1938).

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

[27] Pollaczek, F. Application de la théorie des probabilités à des problèmes posées par l'encombrement des reseaux téléphoniques. *Ann. Télécommunications 14* (1959), 165-183.

[28] Vaulot, E. Delais d'attente des appels téléphoniques traités au hasard. *C.R. Acad. Sci. Paris 222* (1946), 268-269.

[29] Zwart, A.P. Tail asymptotics for the busy period in the $GI/G/1$ queue. *Math. Oper. Res. 26* (2001), 475-483.