Tail asymptotics for processor sharing queues

TU/e technische universiteit eindhoven

SPOR-Report 2003-19

Tail asymptotics for processor sharing queues

F. Guillemin
P. Robert
A.P. Zwart

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven, August 2003
The Netherlands

/department of mathematics and computing science

SPOR-Report
Reports in Statistics, Probability and Operations Research

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

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

ISSN 1567-5211

TAIL ASYMPTOTICS FOR PROCESSOR SHARING QUEUES

FABRICE GUILLEMIN, PHILIPPE ROBERT, AND BERT ZWART

ABSTRACT. The basic queueing system considered in this paper is the $M/G/1$ processor sharing (PS) queue with or without impatience and with finite or infinite capacity. Under some mild assumptions, a criterion for the validity of the RSR (Reduced Service Rate) approximation is established when service times are heavy tailed. This result is applied to various models based on $M/G/1$ processor sharing queues.

CONTENTS

1. INTRODUCTION

Processor Sharing (PS) queues, made popular by the work of Kleinrock [13], were originally proposed to analyze the performance of time sharing disciplines in computer systems. Over the past few years, the processor-sharing paradigm has emerged as a powerful concept for modeling the flow-level performance of bandwidth-sharing protocols in communication networks. In this context, the driving random variables (especially service times) of PS models are often assumed to be heavy-tailed, reflecting the extreme variability of file transfers and session lengths. In view of this, several studies have focused on the analysis of the tail of the sojourn time distribution for the $M/G/1$ PS queue under heavy-tailed assumptions. See for example (in increasing level of generality) Zwart and Boxma [18], Núñez-Queija [14], and Jelenković and Momčilović [10]. In all these references, sufficient conditions are given for the following tail-equivalence between the distributions of the sojourn time V and of the service time B of a customer:

$$\mathbb{P}(V > x) \sim \mathbb{P}(B > x(1-\rho)), \tag{1}$$

The equivalence indicates that for a customer with a large sojourn time everything happens as if he were served alone with a reduced service rate $1-\rho$. In other words, the service rate 1 is reduced by the load ρ of the other customers. Therefore, the equivalence (1) is often called a *reduced-load approximation* (RLA), or also a *reduced service rate approximation* (RSR).

The present paper extends (1) to a wide class of processor sharing queues. This is established by means of a sufficient condition for (1) which can, in principle, be applied to evaluate delay asymptotics for any model in which time-sharing plays a role. In particular, we extend Equation (1) to the case when a customer receives a general service rate f_n if there are n customers in the system (PS is a special case with $f_n = 1/n$). This service discipline is sometimes called Generalized Processor Sharing, see e.g. Cohen [7], but nowadays this term is also used for another class of models (namely Weighted Fair Queueing systems [5]). Thus, we shall call this extension *state-dependent processor sharing.* An important special case of the class of state-dependent PS queues is the $M/G/s$ PS queue, with $f_n = \max\{1, s/n\}$. A convenient property of this class of models is in that the steady-state distribution is *insensitive*; see Bonald and Proutière [3] for recent work on insensitivity of such queues.

A second class of models considered in this paper is composed of PS queues with finite buffers and/or reneging. This class of models is the original motivation for this work: PS queues with reneging aim at capturing impatience effects occurring at the user level. As a matter of fact, it is often observed in the Internet that users transferring large files often interrupt their transactions when the received bandwidth is so small that transaction durations become far too long. This phenomenon is mainly due to the protocol used to transfer data in the Internet (namely TCP), which achieves to some extent a fair sharing of bandwidth in the network..

An important issue for this kind of traffic is whether admission control should be performed or not. The large-deviations results of the present paper aim to give some insight into the introduction of admission control for elastic traffic in broadband packet networks; we refer to the paper Boyer *et al.* [4] for more discussion in this direction. From a technical point of view, the main point of that paper is to make use of the relation $\mathbb{P}(V > x) = \mathbb{P}(B > R(x))$, where $R(x)$ is defined as

$$R(x) = \int_0^x \frac{1}{1 + Q(u)}\, du$$

with $Q(u)$ being the queue length process in the same PS queue *with one additional permanent customer.* In this setting $R(x)$ is the amount of service received by the permanent customer between 0 and x. This relationship holds for any PS queue.

This expression shows the importance of the process $\{R(x)\}$. In fact, the most important condition for (1) to hold is that $\mathbb{P}(R(x) < x/K)$ is sufficiently well-behaved when x gets large for some particular K. This brings us to another contribution of this paper: The study of probabilities of the form $\mathbb{P}(V(ax) > x)$ when $a > 0$ is fixed and x goes to infinity, where $V(ax)$ is the sojourn time of a customer with service time ax. Besides being of intrinsic interest, and being a crucial ingredient in the proof of (1), these asymptotics are also useful to investigate the probability of reneging of large files, see Section 5.

We analyze these probabilities using large-deviations techniques for heavy-tailed distributions. In particular, we show that

$$\mathbb{P}(R(x) < ax) = \mathbb{P}(V(ax) > x) = \mathrm{O}(\mathbb{P}(B^r > x)^{\ell(a)}) \tag{2}$$

with $\ell(a)$ a step function of a. We note that $\ell(a)$ can be interpreted as the number of *other* large customers which are necessary to cause a large sojourn time.

The paper is organized as follows. In Section 2, we formulate our general result. This result is applied in Section 3, to obtain the delay asymptotics for the standard PS queue. The method in this section is easily extended to the class of state-dependent PS queues, as shown in Section 4. Sections 5 and 6 consider PS queues with finite buffers and/or impatience. In particular, an analogue of (1) is developed in Section 5, and analogues of (2) are given in Section 6.

Throughout the paper, for any non-negative random variable X with finite mean, X^r will denote a random variable whose density is given by $\mathbb{P}(X > x)/\mathbb{E}(X)$ and for $x \geq 0$, $\lfloor x \rfloor$ is the integer part of x.

2. A SUFFICIENT CONDITION FOR TAIL-EQUIVALENCE

In this section we give a general result, which will be the starting point for a number of specific models to be discussed later on.

Although all our applications concern PS queues, we will formulate our result in a more general setting. Consider an a.s. increasing stochastic process $(R(x), x \geq 0)$ such that $R(x)/x$ converges to some constant $\gamma \in (0,1)$ a.s., and a random variable B which is independent of $(R(x))$. We are interested in the asymptotic behavior of the probability $\mathbb{P}(B > R(x))$ as x goes to infinity.

The main result of this section gives a sufficient condition for the tail equivalence

$$\mathbb{P}(B > R(x)) \sim \mathbb{P}(B > \gamma x).$$

Informally, this result holds if the speed of convergence of $R(x)/x$ is sufficiently fast, compared to the tail behavior of $\mathbb{P}(B > x)$. This is formalized by Theorem 1 below, which relies on the following conditions.

Assumptions.

(A-1) The distribution function of the service time has a regularly varying tail with index $\nu \geq 1$, i.e.

$$\mathbb{P}(B > x) = L(x)x^{-\nu};$$

(A-2) $R(x)/x \to \gamma$ a.s. as x goes to infinity, with $0 < \gamma < 1$;

(A-3) There exists a positive and finite constant K such that

$$\mathbb{P}(R(x) \leq x/K) = \mathrm{o}(\mathbb{P}(B > x)).$$

Theorem 1. *If Assumptions (A-1)–(A-3) hold and the random variable B is independent of the process $(R(x))$, then*

$$\mathbb{P}(B > R(x)) \sim \mathbb{P}(B > \gamma x)$$

when x goes to infinity.

This result is related to Theorem 5.1.1. in Núñez-Queija [14]. The main strength of our theorem is the weakness of the third condition. In Nunez-Queija's result, a similar condition needs to be satisfied *for every* $K > 1/\gamma$. As indicated in Núñez-Queija [*ibid.*], this condition can be checked when some detailed information of the sojourn-time distribution is known.

Our proof is related to recent work of Foss and Korshunov [9]. Indeed, our result can be reformulated as the problem of sampling the stochastic process $R^{-1}(x)$ at a sub-exponential time B. Further work on this problem can be found in Asmussen *et al.* [2], and Jelenković *et al.* [11].

Proof. For $x \geq 0$ and $\varepsilon > 0$, the quantity $\mathbb{P}(B > R(x))$ can be decomposed as follows:

$$\begin{aligned}\mathbb{P}(B > R(x)) &= \mathbb{P}(B > R(x), R(x) \geq (\gamma+\varepsilon)x) \\ &\quad + \mathbb{P}(B > R(x), (\gamma-\varepsilon)x < R(x) < (\gamma+\varepsilon)x) \\ &\quad + \mathbb{P}(B > R(x), R(x) \leq (\gamma-\varepsilon)x) \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

The three terms are estimated separately. Note that Term I is less than

$$\mathbb{P}(B > (\gamma+\varepsilon)x)\,\mathbb{P}(R(x) \geq (\gamma+\varepsilon)x),$$

which is $o(\mathbb{P}(B > \gamma x))$ by the Law of Large Numbers (LLN) stated above (Assumption (A-2)) and the Regular Variation (RV) property of the distribution of B (Assumption (A-1)).

Term II can be lower bounded by

$$\mathrm{II} \geq \mathbb{P}(\gamma x - \varepsilon x < R(x) < \gamma x + \varepsilon x)\,\mathbb{P}(B > \gamma x + \varepsilon x) = \mathbb{P}(B > \gamma x + \varepsilon x)\,(1 - o(1)),$$

as $x \to \infty$, and upper bounded by $\mathrm{II} \leq \mathbb{P}(B > \gamma x - \varepsilon x)$.

We now turn to Term III. Distinguishing between the two cases $R(x) < x/K$ and $R(x) \geq x/K$, the quantity III can be written as III = IIIa+IIIb. Assumption (A-3) implies that IIIa can be neglected. The other term IIIb is given by

$$\begin{aligned}\mathrm{IIIb} &= \mathbb{E}\left[\mathbb{P}(B > R(x)|R(x))\,1_{\{x/K<R(x)<(\gamma-\varepsilon)x\}}\right] \\ &\leq \mathbb{P}(B > x/K)\,\mathbb{P}(R(x) < (\gamma-\varepsilon)x) \\ &= \mathbb{P}(B > \gamma x)\,\frac{\mathbb{P}(B > x/K)}{\mathbb{P}(B > \gamma x)}\mathbb{P}(R(x) < (\gamma-\varepsilon)x).\end{aligned}$$

The regular variation property of the distribution of B shows that second term of the last expression is converging to a constant as x goes to infinity. By Assumption (A-2), it is then clear that IIIb is $o(\mathbb{P}(B > \gamma x))$.

As a consequence, since

$$\frac{\mathbb{P}(B > \gamma x \pm \varepsilon x)}{\mathbb{P}(B > \gamma x)} = \left(\frac{\gamma}{\gamma \pm \varepsilon}\right)^{\nu}(1 + o(1))$$

when x goes to infinity, we deduce that for all $\varepsilon > 0$

$$\left(\frac{\gamma}{\gamma+\varepsilon}\right)^{\nu} + o(1) \leq \frac{\mathbb{P}(B > R(x))}{\mathbb{P}(B > \gamma x)} \leq \left(\frac{\gamma}{\gamma-\varepsilon}\right)^{\nu} + o(1)$$

when x goes to infinity. Since the above inequality is valid for all $0 < \varepsilon < \gamma$, we obtain the desired result and the theorem is proved. □

As will become clear below, Assumption (A-3) is the most difficult condition to be checked. Assumption (A-2) is usually implied by stability of the queueing system. Indeed, when

$$R(x) = \int_0^x \frac{1}{1+Q(u)}\,du,$$

where $(Q(u))$ is the number of non-permanent customers in a processor sharing queue having a permanent customer, $R(x)$ is the amount of service received by the permanent customer. In this case, under some stability condition,

$$\gamma = \mathbb{E}\left(\frac{1}{1+Q(\infty)}\right) \tag{3}$$

where $Q(\infty)$ is the limit in distribution of the process $(Q(u))$.

3. The $M/G/1$ PS queue

In this section we consider the standard $M/G/1$ PS queue with arrival rate λ and generic service time distribution B. Let $\rho = \lambda\mathbb{E}(B)$ be the load of the queue, assumed to be strictly less than 1. The goal of this section is to show that Theorem 1 can be used in this case. While the reduced load approximation (RLA) is already known for this queue in a more general context, this section is intended to set up a procedure relying on the use of Assumptions (A-1)–(A-3), which will used in the following to obtain an RLA property for several other queueing systems.

The processor sharing queue is assumed to be at equilibrium at time 0 and that a permanent customer arrives at that time. The stationary sojourn time V of this queue can then be expressed as the time at which the amount of service received by the permanent customer is equal to his requested service time B independent of the system. This yields the following representation: $\mathbb{P}(V > x) = \mathbb{P}(B > R(x))$, where

$$R(x) = \int_0^x \frac{1}{1+Q(u)}\,du,$$

$(Q(u))$ describing the number of non-permanent customers in the queue. It is first shown that Assumption (A-2) holds for the variables $(R(x))$. By adding residual service times of non-permanent customers, it is not difficult to show that the process $(Q(t))$ can be embedded in a Markov process and that this process is Harris ergodic (by looking at the evolution of the total workload and by using Proposition 3.13 of Asmussen [1] for example). Thus the ergodic Theorem for Harris Markov chains shows that $(x - R(x))/x$ converges almost surely to some constant. Since $x - R(x)$ is the amount of work received by non-permanent customers up to time x, this constant is necessarily ρ. Hence the quantity $R(x)/x$ converges almost surely to $1-\rho$ as x goes to infinity. Assumption (A-2) is hence satisfied.

A key ingredient in our analysis is the Processor Sharing queue with k permanent customers, the corresponding queue length process is denoted by $Q_k(u)$. Define

$$R_k(x) = \int_0^x \frac{1}{k+Q_k(u)}\,du.$$

Note that by definition $R(x) = R_1(x)$. With similar arguments as before, it is not difficult to show that

$$\lim_{x \to +\infty} \frac{R_k(x)}{x} = \frac{1}{k}\mathbb{E}\left(1 - \frac{Q_k(\infty)}{k + Q_k(\infty)}\right) = \frac{1-\rho}{k}. \tag{4}$$

The main result of this section is the following RLA property.

Theorem 2. *If $\rho < 1$ and B is regularly varying of index ν with $\nu > 1$, then*

$$\mathbb{P}(V > x) \sim \mathbb{P}(B > (1-\rho)x),$$

as x goes to infinity.

In order to prove Theorem 2, the major difficulty is to show that Assumption (A-3) holds. For, we prove a series of technical lemmas. But before proceeding with this task, let us introduce some additional notation.

Let $C(\varepsilon, x)$ be the number of (non-tagged) customers in the system with service time larger than εx. More precisely, $C(\varepsilon, x)$ counts both those customers in the system at time 0 with remaining service time exceeding εx and those customers entering the system in the time interval $(0, x)$ with service time larger than εx. Moreover, we add a subscript "$<\varepsilon x$", if we consider a system where all service times are conditioned to be smaller than εx, including those customers already in the queue at time 0. Finally, let $\ell(a)$ be defined as:

$$\ell(a) = \inf\left\{k : \frac{(1-\rho)}{(k+1)} < a\right\} = \left\lfloor \frac{(1-\rho)}{a} \right\rfloor. \tag{5}$$

We first state an important lemma, which is an easy consequence of equation (3.9) in Zwart [17].

Lemma 1. *Consider an $M/G/1$ FIFO queue with input rate λ and such that service times are bounded by εx for some $\varepsilon > 0$ and $x > 0$. Let $P(\varepsilon x)$ denote the duration of a busy period of this queue starting at time $t = 0$. Then, for every $\beta > 0$, there exists $\varepsilon_0 > 0$ such that for $\varepsilon < \varepsilon_0$*

$$\mathbb{P}(P(\varepsilon x) > x) = o(x^{-\beta}),$$

when x goes to infinity.

By using the above lemma, we can then prove the following result.

Lemma 2. *Let $0 < a < 1$ and suppose that $(1-\rho)/a$ is not an integer. Then, for every $\beta > 0$, there exists $\varepsilon > 0$ such that*

$$\mathbb{P}(R(x) < ax, C(\varepsilon, x) \leq \ell(a) - 1) = \mathrm{o}\left(x^{-\beta}\right).$$

Proof. Let $\beta > 0$ and consider first the case $\ell(a) = 1$. The probability of interest has the form

$$\mathbb{P}(R(x) < ax, C(\varepsilon, x) = 0) = \mathbb{P}(V_{<\varepsilon x}(ax) > x),$$

where $V_{<\varepsilon x}(ax)$ is the sojourn time of a customer with requested service time ax under the assumption that all other service times are smaller than εx.

Since the standard PS queue is work conserving, we can bound $V_{<\varepsilon x}(ax)$ by the residual busy period of an $M/G/1$ queue with service times $B < \varepsilon x$ and an additional customer with service time ax. Write this residual busy period as $P(a, \varepsilon, x)$. Without loss of generality, we can first serve the customer with service time ax, and all the work arrived during this first service (and so on).

This branching argument leads to the decomposition

$$P(a, \varepsilon, x) = \overline{P}_1(a, \varepsilon, x) + \overline{P}_2(\varepsilon, x),$$

where $\overline{P}_1(a, \varepsilon, x)$ is the busy period initiated by the customer with service time ax and $\overline{P}_2(\varepsilon, x)$ is the busy period initiated by customers in the queue at time 0. We then have

$$\begin{aligned}\mathbb{P}\left(V_{<\varepsilon x}(ax) > x\right) &\leq \mathbb{P}\left(P(a, \varepsilon, x) > x\right) \\ &\leq \mathbb{P}\left(\overline{P}_1(a, \varepsilon, x) > \delta x\right) + \mathbb{P}\left(\overline{P}_2(\varepsilon, x) > (1-\delta)x\right),\end{aligned}$$

for some $\delta \in (a, 1)$.

Let $P(\varepsilon x)$ be the duration of a busy period of an $M/G/1$ queue with service times less than εx. Moreover, let $M(ax)$ be the number of customers entering the system during the service of the large customer. Note that

$$\overline{P}_1(a, \varepsilon, x) \overset{\text{dist.}}{=} ax + \sum_{i=1}^{M(ax)} P_i(\varepsilon x).$$

where $P_i(\varepsilon x)$ are independent copies of $P(\varepsilon x)$. Hence,

$$\begin{aligned}\mathbb{P}\left(\overline{P}_1(a, \varepsilon, x) > \delta x\right) &= \mathbb{P}\left(\sum_{i=1}^{M(ax)} P_i(< \varepsilon x) > (\delta - a)x\right) \\ &\leq \mathbb{P}\left(M(ax) \geq \lfloor \lambda a + \eta)x \rfloor\right) + \mathbb{P}\left(\sum_{i=1}^{\lfloor \lambda a + \eta)x \rfloor} P_i(\varepsilon x) > (\delta - a)x\right).\end{aligned}$$

for some $\eta > 0$. The second term on the r.h.s. of the above equation can be upper bounded by

$$\lfloor \lambda a + \eta)x \rfloor \mathbb{P}\left(P_{<\varepsilon x} > qx\right) + \mathbb{P}\left(\sum_{i=1}^{\lfloor \lambda a + \eta)x \rfloor} P_i \wedge (qx) > (\delta - a)x\right)$$

for some $q > 0$, where P_i are i.i.d. copies of the busy period P of the $M/G/1$ FIFO queue with service times regularly varying with index ν and where we have used the fact that $P(\varepsilon x) \leq_{st} P$. Note that the mean value of P is $\rho/(\lambda(1-\rho))$.

In the following, we show that by adequately choosing the parameters q, η, δ and ε, we have $\mathbb{P}\left(\overline{P}_1(a, \varepsilon, x) > \delta x\right) = o(x^{-\beta})$ as x gets large. Since it has been assumed that $\ell(a) = 1$, in particular $a/(1-\rho) < 1$, a constant $\delta \in (a/(1-\rho), 1)$ is fixed. Choose now $\eta > 0$ and $\mu > 0$ so that

$$\kappa \overset{\text{def.}}{=} \delta - (\lambda a + \eta)\left(\frac{1}{\lambda(1-\rho)} + \mu\right) > 0.$$

With this choice of the parameters, we have

$$\mathbb{P}\left(\sum_{i=1}^{\lfloor(\lambda a+\eta)x\rfloor} P_i \wedge (qx) > (\delta - a)x\right)$$
$$\leq \mathbb{P}\left(\sum_{i=1}^{\lfloor(\lambda a+\eta)x\rfloor} P_i \wedge (qx) > \lfloor(\lambda a+\eta)x\rfloor\left(\frac{\rho}{\lambda(1-\rho)}+\mu\right)+\kappa x\right)$$

Since we know from Zwart [17] that the tail distribution of the variable P is dominated at infinity by a regularly varying function with index ν, by applying Lemma 2.1 in Resnick and Samorodnitsky [15], under the assumption $\lfloor\kappa/q\rfloor \geq 3$, one gets that

$$\mathbb{P}\left(\sum_{i=1}^{\lfloor(\lambda a+\eta)x\rfloor} P_i \wedge (qx) > (\delta - a)x\right) \leq \varphi(\kappa x),$$

where φ is regularly varying with index $\lfloor\lfloor\kappa/q\rfloor/3\rfloor(\nu-1)$ at infinity. By choosing q sufficiently small so that $\lfloor\lfloor\kappa/q\rfloor/3\rfloor(\nu-1) > \beta$, we deduce that when x gets large,

$$\mathbb{P}\left(\sum_{i=1}^{\lfloor(\lambda a+\eta)x\rfloor} P_i \wedge (qx) > (\delta - a)x\right) = o(x^{-\beta}).$$

By choosing ε sufficiently small, the relation $\mathbb{P}(P(\varepsilon x) > qx) = o(x^{-\beta-1})$ holds as x goes to infinity by Lemma 1, thus

$$(\lambda a+\eta)x\mathbb{P}(P_{<\varepsilon x} > qx) = o(x^{-\beta})$$

when x gets large. Finally, the term $\mathbb{P}(M(ax) > \lfloor(\lambda a+\eta)x\rfloor)$ decays exponentially fast and is then $o(x^{-\beta})$. The above results show that $\mathbb{P}\left(\overline{P}_1(a,\varepsilon,x) > \delta x\right) = o(x^{-\beta})$ as x gets large.

To complete the proof for $\ell(a) = 1$, we note that we simply have

$$\mathbb{P}\left(\overline{P}_2(\varepsilon,x) > (1-\delta)x\right) = \mathbb{P}\left(P^r_{<\varepsilon x} > (1-\delta)x\right),$$

where $P^r_{<\varepsilon x}$ is the residual lifetime of the busy period $P_{<\varepsilon x}$. Since we can choose ε sufficiently small so that $\mathbb{P}(P_{<\varepsilon x} > x) = o(x^{-(\beta+1)})$ when $x \to \infty$, hence

$$\mathbb{P}(P^r(\varepsilon x) > (1-\delta)x) = o(x^{-\beta}).$$

Let us now consider the case when $\ell(a)$ is arbitrary. We have

$$\mathbb{P}(R(x) < ax, C(\epsilon,k) \leq \ell(a)-1) \leq \mathbb{P}\left(R_{\ell(a),<\epsilon x} < ax\right),$$

with

$$R_{\ell(a),<\epsilon x} = \int_0^x \frac{1}{\ell(a) + Q_{\ell(a),<\epsilon x}(u)}\,du.$$

Note that the latter probability is related to the sojourn time distribution in a PS queue with $\ell(a)$ permanent customers:

$$\mathbb{P}\left(R_{\ell(a),<\epsilon x} < ax\right) = \mathbb{P}\left(V_{\ell(a),<\epsilon x}(ax) > x\right).$$

It has been shown by van den Berg and Boxma [16] that the Laplace transform of the sojourn time $\mathcal{V}(\tau)$ of a customer with service time τ in the $M/G/1$ PS queue conditionally to the fact that there are k permanent customers is given by

$$\mathbb{E}\left(e^{-s\mathcal{V}(\tau)} \mid k,\tau\right) = \left(\mathbb{E}\left(e^{-s\mathcal{V}(\tau)} \mid 0,\tau\right)\right)^{k+1}. \tag{6}$$

Taking $\tau = \infty$, we deduce that the sojourn time $V_{\ell(a),<\varepsilon x}(ax)$ is such that

$$V_{\ell(a),<\varepsilon x}(ax) \overset{\text{dist.}}{=} V^1_{<\varepsilon x}(ax) + \cdots + V^{\ell(a)}_{<\varepsilon x}(ax),$$

where $V^i_{<\varepsilon x}(ax)$ for $i = 1,\dots,\ell(a)$ are $\ell(a)$ independent copies of $V_{<\varepsilon x}(ax)$, which is the sojourn time of the customer with service time ax. It follows by using the union bound that

$$\begin{aligned}\mathbb{P}\left(V_{\ell(a),<\varepsilon x}(ax) > x\right) &\leq \ell(a)\mathbb{P}\left(V_{<\varepsilon x}(ax) > x/\ell(a)\right) \\ &= \ell(a)\mathbb{P}\left(R_{<\varepsilon x}(x/\ell(a)) < ax\right).\end{aligned}$$

We can then proceed as in the case $\ell(a) = 1$, using the fact that $a < (1-\rho)/\ell(a)$, and we get the desired result. This completes the proof. □

Now, we investigate the asymptotic behavior of the quantity $\mathbb{P}(R(x) < ax)$.

Lemma 3. *Suppose that $a < 1-\rho$ and that $(1-\rho)/a$ is not an integer. Then,*

$$\mathbb{P}(R(x) < ax) = O\left((x\mathbb{P}(B > x))^{\ell(a)}\right), \tag{7}$$

when x goes to infinity, where $\ell(a)$ is defined by Equation (5).

Proof. Let $\beta > 0$. Using Lemma 2, we have for suitably small ε

$$\mathbb{P}(R(x) < ax) = \mathrm{o}\left(x^{-\beta}\right) + \mathbb{P}\left(R(x) < ax, C(\varepsilon,x) \geq \ell(a)\right).$$

The latter probability is smaller than $\mathbb{P}(C(\varepsilon,x) \geq \ell(a))$. Note that

$$C(\varepsilon,x) \leq C_0 + C_{(0,x]},$$

where C_0 is the number of large customers present in the system at time 0, and $C_{(0,x]}$ is the number of arrivals with service time larger than εx. The variable $C_{(0,x]}$ has a Poisson distribution with rate $\lambda x\mathbb{P}(B > \varepsilon x)$ and C_0 has a geometric distribution with parameter $\rho\mathbb{P}(B^r > \varepsilon x)/(1-\rho+\rho\mathbb{P}(B^r > \varepsilon x))$, i.e. for $m \geq 0$,

$$\mathbb{P}(C_0 \geq m) = \left(\frac{\rho\mathbb{P}(B^r > \varepsilon x)}{1-\rho+\rho\mathbb{P}(B^r > \varepsilon x)}\right)^m,$$

where B^r is the random variable with density function $\mathbb{P}(B \geq x))/\mathbb{E}(B)$ on $\mathbb{R}_+$. Hence,

$$\begin{aligned}\mathbb{P}(C(\varepsilon,x) \geq \ell(a)) &= \mathbb{P}\left(C_{(0,x]} \geq \ell(a)\right) \\ &+ e^{-\lambda x\mathbb{P}(B>\varepsilon x)} \sum_{i=0}^{\ell(a)-1} \frac{(\lambda x\mathbb{P}(B > \varepsilon x))^i}{i!} \left(\frac{\rho\mathbb{P}(B > \varepsilon x)}{1-\rho+\rho\mathbb{P}(B^r > \varepsilon x)}\right)^{\ell(a)-i}.\end{aligned}$$

Using the fact that $\mathbb{P}(B^r > x) = O(x\mathbb{P}(B > \varepsilon x))$ when $x \to \infty$, it is easy to see that all terms in the summation are $O((x\mathbb{P}(B > x))^{\ell(a)})$. The first term can be upper bounded by $(\lambda x\mathbb{P}(B > \varepsilon x))^{\ell(a)}/\ell(a)!$, since the distribution of $C_{(0,x]}$ is

Poisson. Taking β sufficiently large (e.g., $\beta > (\nu - 1)\ell(a)$), we obtain the desired result. This completes the proof. □

We are now ready to prove Theorem 2.

Proof of Theorem 2. It is sufficient to check that Assumption (A-3) holds for the model under consideration. For, we use Lemma 3. Take a small enough such that $\ell(a)(\nu - 1) > \nu$. Then

$$\mathbb{P}\left(R(x) < ax\right) = \mathrm{o}(\mathbb{P}\left(B > x\right)).$$

Assumption (A-3) then holds with $K = 1/a$. The theorem is proved. □

To conclude this section, let us give a heuristic explanation of Lemma 3. In order for the event $\{R(x) < ax\}$ to take place, the drift of the process $R(x)$ should be smaller than a for a long time. In average, the drift of $R(x)$ is $1 - \rho$. Thus, if $a < 1 - \rho$ one or more rare events to let the drift decrease below a are required.

It turns out that the most likely way for this change of drift to occur, is the presence of other large customers. These customers can be regarded as additional (besides the customer under consideration) permanent customers. For a model with $k + 1$ permanent customers, Equation (4) shows that, to make the limit of $R(x)/x$ smaller than a, at least $\ell(a)$ other customers with service times of the order of $\mathrm{O}(x)$ are needed. Hence, the estimate (7) should hold.

4. State dependent PS queues

In this section the reduced load equivalence is extended to the class of state dependent PS disciplines as investigated in Cohen [7], Kelly [12] and Bonald and Proutière [3]. As mentioned in the Introduction, this model is an extension of the standard PS queue, in the sense that each customer is being served at rate f_n instead of $1/n$ if there are n customers in the system. In the following, the sequence $f = (f_n, n \geq 1)$ is fixed and such that $f_n > 0$ for each $n > 0$. Note that, if there is a permanent customer, this amounts to consider a state dependent PS queue associated to the shifted sequence (f_{n+1}). The quantity ϕ_n is defined by

$$\phi_n = \frac{1}{n! \prod_{i=2}^{n+1} f_i},$$

with the convention that $\phi_0 = 1$. An important special case is the $M/G/s$ PS queue for which $f_n = \max(1, s/n)$.

A superscript f is added to indicate the dependence on the sequence f. As before a permanent customer is assumed to arrive in the queue at time 0 and $V^f(x)$ is the time for this customer to receive service x. The amount of service $R^f(x)$ received by the permanent customer at time x is given by

$$R^f(x) = \int_0^x f_{1+Q^f(u)}\, du,$$

where Q^f refers to the number of non-permanent customers.

Assumption.

(A-4) It is assumed that

$$\liminf_{n\to+\infty} n f_{n+1} > \rho = \lambda \mathbb{E}(B).$$

With this assumption, the queue is stable. The main result of this section is the following theorem.

Theorem 3. *If the service time distribution is regularly varying with index $\nu \geq 1$, and if Assumption (A-4) holds then*

$$\mathbb{P}(V^f > x) \sim \mathbb{P}(B > \gamma^f x),$$

with

$$\gamma^f = \sum_{n=0}^{\infty} \rho^n \phi_n f_{n+1} \Big/ \sum_{n=0}^{\infty} \rho^n \phi_n . \tag{8}$$

Note that by Cohen [7], $\mathbb{E}(V^f(x)) = x/\gamma^f$. Theorem 3 is proved by checking Assumptions (A-2) and (A-3) (see Proposition 1 and Proposition 4 below) and by using Theorem 1.

Proposition 1. *If Assumption (A-4) holds, then, almost surely,*

$$\lim_{x\to+\infty} \frac{R^f(x)}{x} = \gamma^f,$$

with γ^f defined by Equation (8).

Proof. Let $(M(t))$ is a Poisson process with parameter λ. The process $W^f(t)$ describing the workload in the queue at time t is governed by the evolution equation:

$$W^f(t) = W^f(0) + \sum_{i=1}^{M(t)} B_i - \int_0^t Q^f(s) f_{1+Q^f(u)}\, du,$$

where (B_i) is an i.i.d. sequence with the same distribution as B. For $K > 0$, T_K denotes the first time the process $(Q(t))$ is in the interval $[0, K]$, then Wald's Formula gives

$$\mathbb{E}\left(W^f(t\wedge T_K)\right) = \mathbb{E}\left(W^f(0)\right) + \lambda\mathbb{E}(B)\,\mathbb{E}(t\wedge T_K) - \mathbb{E}\left(\int_0^{t\wedge T_K} Q^f(u) f_{Q^f(u)+1} du\right).$$

Assumption (A-4) shows that there exist $\varepsilon > 0$ and K_0 such that $n f_{n+1} > \rho + \varepsilon$ for $n \geq K_0$. The above identity gives the relationship

$$0 \leq \mathbb{E}\left(W^f(0)\right) - \varepsilon\mathbb{E}(t\wedge T_{K_0}),$$

by letting t go to infinity, one gets that $\mathbb{E}(T_{K_0}) < \mathbb{E}\left(W^f(0)\right)/\varepsilon$. The variable T_{K_0} is integrable.

By adding residual service times the process $(Q^f(t))$ becomes a Markov process and it is then not difficult to conclude then that the Markov process associated with $(Q^f(t))$ is Harris ergodic. Ergodic Theorem for Harris Markov chains gives the almost sure convergence

$$\lim_{x\to+\infty} \frac{R^f(x)}{x} = \mathbb{E}\left(f_{1+Q^f(\infty)}\right).$$

Since, for $n \geq 1$, $\mathbb{P}\left(Q^f(\infty) = n\right) = \rho^n\phi_n$, the last expression is indeed γ^f. □

The above result shows that Assumption (A-2) is valid. To show that Assumption (A-3) holds, the process $(R^f(x))$ will be lower-bounded by the process associated with a PS queue with k permanent customers, for some convenient $k \geq 1$. The following lemma establishes a monotonicity property of the mapping $f \to (R^f(x))$.

Lemma 4. *If, for some non-negative sequences $f = (f_n)$ and $g = (g_n)$, the following conditions are satisfied:*

- *For any $n \geq 1$, $f_n \geq g_n$;*
- *The sequence g is non-increasing,*

then, for any $a \geq 0$ and $x \geq 0$ and any initial state of the queue, the relation $\mathbb{P}\left(R^f(x) \geq a\right) \leq \mathbb{P}\left(R^g(x) \geq a\right)$ holds. In other words, the variable $R^f(x)$ is stochastically dominated by $R^g(x)$ that is, $R^f(x) \leq_{st} R^g(x)$.

Proof. To compare the two queues, a sample path argument is used. Both queues have the same arrival process and same services. Before the first departure in one of the queues, a job in the f-queue (i.e. the PS queue with the sequence f) receives the service $f_{1+Q^f(t)}\,dt \geq g_{1+Q^f(t)}\,dt = g_{1+Q^g(t)}\,dt$ during time interval $[t, t+dt]$. Consequently, the first departure will occur in the f-queue, and a that time a customer present in the f-queue is also in the g-queue with a larger residual service time. Thus, if t is smaller that the second departure time from the f-queue, then $Q^f(s) \leq Q^g(s)$ for $s \leq t$ and

$$\int_0^t f_{1+Q^f(u)}\,du \geq \int_0^t g_{1+Q^f(u)}\,du \geq \int_0^t g_{1+Q^g(u)}\,du,$$

since the sequence g is non-increasing. Thus, these inequalities also hold until the third departure. By induction on the number of departures from the f-queue, one concludes that $Q^f(t) \leq Q^g(t)$ for any $t \geq 0$, and therefore that $R^f(x) \geq R^g(x)$ for any $x \geq 0$. The lemma is proved. □

The main result of this section can now be proved.

Proof of Theorem 3. By Proposition 1 and Theorem 1, only Assumption (A-3) has to be checked. Assumption (A-4) gives an $\varepsilon > 0$ and $n_0 \geq 1$ such that $nf_n \geq \rho+\varepsilon$ for $n \geq n_0$. Therefore, if k is chosen sufficiently large, the inequality $(n+k)f_n \geq \rho+\varepsilon$ holds for any $n \geq 1$. Hence of $g_n = (\rho+\varepsilon)/(n+k)$ for $n \geq$, the sequence $g = (g_n)$ corresponds to a PS queue with k permanent customers and service speed $\rho+\varepsilon$. As in the proof of Proposition 2, the convolution result (6) gives the inequalities

$$\mathbb{P}\left(R^f(x) < x/K\right) \leq \mathbb{P}\left(R^g(x) < x/K\right) \leq k\mathbb{P}\left(R(x/k) < x/K\right),$$

for $K \geq 0$. Hence, the proposition follows from Theorem 3. □

5. PS QUEUES WITH BLOCKING AND/OR RENEGING

In this section, we introduce a new extension of the standard $M/G/1$ PS queue. It is specifically assumed that the total number of customers is limited by some threshold $N > 0$, and that the sojourn time of a customer cannot exceed some value I. In other words, when the sojourn time exceeds I, the customer renegs by leaving the system. The variables B and I can a priori be dependent (e.g., $I = \theta B$).

In spite of its obvious practical relevance (see for instance [4] for the connection with the problem of admission control in packet networks), little is known about PS queues with blocking and reneging. An exception is the paper by Coffmann *et al.* [6], where B and I are independent and both exponentially distributed. The main goal of the present section is to determine the delay asymptotics of a non-reneging customer. In Section 6, the probability of reneging for large customers is investigated in the special case $I = \theta B$.

Denote the sojourn time of a non-reneging customer by $\overline{V}$. As in previous sections, the tail behavior of the distribution of $\overline{V}$ is given by

$$\mathbb{P}\left(\overline{V} > x\right) = \mathbb{P}\left(B > R(x)\right),$$

with again $R(x) = \int_0^x 1/(1+Q(u))\,du$. Here, $Q(u)$ is the number of customers in a PS queue with finite waiting room N (i.e., there is room for $N-1$ non-permanent customers) and reneging. The original sojourn time V can be written as

$$V = \min\{\overline{V}, I\},$$

and, with obvious notation,

$$V(x) = \min\{\overline{V}(x), I(x)\}.$$

Note that $\overline{V}$ and I are dependent in general.

It is clear that the process $Q(u)$ is ergodic if $N < \infty$ or $I < \infty$ a.s. Hence, Assumption (A-2) is always satisfied, and we have a.s.

$$R(x)/x \to \gamma$$

as x goes to infinity.

We now verify that Assumptions (A-2) and (A-3) are satisfied under weaker conditions. The next proposition guarantees the validity of Assumption (A-2).

Proposition 2. *If* $\mathbb{E}\,(I) < \infty$ *or* $N < \infty$ *or* $\rho < 1$, *then there exists a constant* $\gamma = \gamma_{N,I} \in (0,1)$ *such that, almost surely,*

$$R(x)/x \to \gamma.$$

Proof. If $\rho < 1$, then one can upper bound $Q(u)$ by letting $N = I = \infty$, so that the number of customers of this queue is stochastically smaller than the number of customers in a stable $M/G/1$ queue (i.e. the queue without reneging).

If $\mathbb{E}\,(I) < \infty$, the busy period of this queue is stochastically smaller than the busy period of an $M/G/\infty$ queue with service times I. In particular it has finite mean if $\mathbb{E}\,(I) < \infty$.

In both cases the number of customers of this queue can thus be embedded into an Harris ergodic Markov chain. The convergence result of the proposition is just an ergodic theorem for the Markov chain. The result is trivial if $N < \infty$. □

Theorem 4. *Let B be regularly varying of index $\nu > 1$. If $N < \infty$ or $\rho < 1$ or there exists some $p > 1$ such that $\mathbb{E}(I^p) < \infty$, then*

$$\mathbb{P}\left(\overline{V} > x\right) \sim \mathbb{P}\left(B > x\gamma_{N,I}\right)$$

when x goes to infinity, where $\gamma_{N,I}$ is defined in Proposition 2.

In general, it is not possible to compute $\gamma_{N,I}$. An exception is the case $I \equiv \infty$. It is easily seen that in this case, the model is equivalent to one of the queues presented in Section 4 with $f_n = \infty$ when $n > N$. From this, it follows easily that

$$\gamma_N \overset{def}{=} \gamma_{N,\infty} = (1-\rho)\frac{1}{1 - N\rho^N(1-\rho)/(1-\rho^N)}. \tag{9}$$

Proof of Theorem 4. It suffices to show that Assumption (A-3) is satisfied. This is trivial if $N < \infty$, because this implies $R(x) \geq x/N$. If $\rho < 1$, then one can lower bound $R(x)$ by setting $N = I = \infty$ and apply Theorem 3 for the ordinary $M/G/1$ PS queue.

Suppose that $\mathbb{E}(I^p) < \infty$. The main idea is to lower-bound $R(x)$ by bounding $Q(u)$ with the queue length process of the $M/G/\infty$ queue with service times I, i.e., we assume that all (non-permanent) customers remain in the system until they reneg. Let $(Q_1(u))$ be the process of the number of customers in this $M/G/\infty$ queue.

From the above argument, we obtain $Q(u) \leq Q_1(u)$. Hence,

$$R(x) \geq \int_0^x \frac{1}{1+Q_1(u)}\,du.$$

The idea is to write the integral as a sum, using the regenerative structure of the $M/G/\infty$ queue. The number of summands then corresponds to the number of cycles of the $M/G/\infty$ queue.

As before, for $\varepsilon > 0$, a customer is said to be small if its service time is smaller than εx with ε some small constant, to be chosen later on. Otherwise, the customer is said to be large.

Let $H_1(\varepsilon, x)$ be the number of large customers in the system at time 0. Since the evolution of the number of large customers can be viewed as a separate $M/G/\infty$ with arrival rate $\lambda\mathbb{P}(I > \varepsilon x)$ and mean service times $\mathbb{E}(I \mid I > \varepsilon x)$ it follows that $H_1(\varepsilon x)$ has a Poisson distribution with rate $\lambda\mathbb{E}\left(I1_{\{I>\varepsilon x\}}\right)$.

Next, we define $H_2(\varepsilon, x)$ as the number of large customers entering the system between time 0 and x. Obviously, $H_2(\varepsilon, x)$ has a Poisson distribution with mean $\lambda x\mathbb{P}(I > \varepsilon x)$.

Finally, set $H(\varepsilon, x) = H_1(\varepsilon, x) + H_2(\varepsilon, x)$. Note that $H(\varepsilon, x)$ has a Poisson distribution with rate

$$\mu_1(\varepsilon x) \overset{def}{=} \lambda\mathbb{E}\left(I(I > \varepsilon x)\right) + \lambda x\mathbb{P}\left(I > \varepsilon x\right). \tag{10}$$

Since I has a finite pth moment, Chebyshev's inequality gives some constant C (which depends upon ε) such that

$$\mu_1(\varepsilon x) \leq C x^{1-p}.$$

Since $H(\varepsilon, x)$ has a Poisson distribution, for each k and each $\varepsilon > 0$, one has

$$\mathbb{P}\left(H(\varepsilon,x) > k\right) = \mathrm{O}\left(x^{-k(p-1)}\right) = \mathrm{o}(\mathbb{P}\left(B > x\right)), \tag{11}$$

if k is chosen sufficiently large so that $k(p-1) > \nu$. From now on, we choose k such that this inequality is satisfied. Therefore,

$$\begin{aligned}\mathbb{P}\left(R(x) \leq x/K\right) &\leq \mathbb{P}\left(\int_0^x \frac{1}{1+Q_1(u)}\,du \leq x/K\right)\\ &= \mathrm{I} + \mathrm{II},\end{aligned}$$

where we separate between the possibilities $H(\varepsilon, x) \leq k$ and $H(\varepsilon, x) > k$.

We need to show that both terms are of $\mathrm{o}(\mathbb{P}\left(B > x\right))$. For the second term, note that

$$\mathrm{II} \leq \mathbb{P}\left(H(\varepsilon,x) > k\right) = \mathrm{O}\left(x^{-k(p-1)}\right) = \mathrm{o}(\mathbb{P}\left(B > x\right)).$$

To upper bound the first term, note that if $H(\varepsilon, x) \leq k$,

$$Q_1(u) \leq k + Q_{1,<\varepsilon x}(u),$$

with $Q_{1,<\varepsilon x}(u)$ denoting the number of customers at time u in an $M/G/\infty$ queue with all service times smaller than εx. For this particular $M/G/\infty$ queue, we define $\tau(\varepsilon, x)$ as the number of busy periods completed at time x. Note that, during each busy period $i \geq 1$, our permanent customer gets at least $E_i/(k+1)$ units of service, where E_i is an exponential random variable with mean $1/\lambda$.

Hence, we can conclude that, if $H(\varepsilon, x) \leq k$,

$$R(x) \geq \frac{1}{k+1} \sum_{i=1}^{\tau(\varepsilon,x)} E_i. \tag{12}$$

Thus, it suffices to show that there exists a finite constant K such that

$$\mathbb{P}\left(\sum_{i=1}^{\tau(\varepsilon,x)} E_i \leq x/K\right) = \mathrm{o}(\mathbb{P}\left(B > x\right)). \tag{13}$$

Write, for some $a > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{\tau(\varepsilon,x)} E_i \leq x/K\right) \leq \mathbb{P}\left(\tau(\varepsilon,x) < \lfloor ax \rfloor\right) + \mathbb{P}\left(\sum_{i=1}^{ax} E_i \leq x/K\right).$$

The second term clearly decreases exponentially fast in x, as long as we choose K such that $a\lambda > 1/K$. Thus, it remains to show that, for some ε and a,

$$\mathbb{P}\left(\tau(\varepsilon,x) < ax\right) = \mathrm{o}(\mathbb{P}\left(B > x\right)).$$

For $\delta > 0$,

$$\begin{aligned}\mathbb{P}\left(\tau(\varepsilon, x) < ax\right) &\leq \mathbb{P}\left(\pi^0(\varepsilon x) + \pi^1(\varepsilon x) + \cdots + \pi^{\lfloor ax \rfloor}(\varepsilon x) > x\right) \\ &\leq \mathbb{P}\left(\pi^0(\varepsilon x) > \delta x\right) + \mathbb{P}\left(\pi^1(\varepsilon x) + \cdots + \pi^{\lfloor ax \rfloor}(\varepsilon x) > (1-\delta)x\right) \\ &= \mathrm{III} + \mathrm{IV}.\end{aligned}$$

In the above expression, $\pi^i(\varepsilon x)$ denotes the i-th busy cycle of the $M/G/\infty$ queue with service times smaller than εx, and $\pi^0(\varepsilon x)$ the remaining busy cycle at time 0. It is easy to see that $\pi^0(\varepsilon x) \overset{d}{=} \pi^1(\varepsilon x)^r$.

Note that $\mathbb{E}\left(\pi^1(\varepsilon x)\right) \leq p_1 := e^{\lambda \mathbb{E}(I)}$. Choose a and δ small enough such that $ap_1 < (1-\delta)$. Proposition 1 of Resnick and Samorodnitsky [15] shows that for each $\varepsilon > 0$ there exists a $\beta > 0$ such that

$$\mathbb{P}\left(\pi^1(\varepsilon x) > x\right) = \mathrm{o}\left(x^{-\beta}\right)$$

and therefore, the relation

$$\mathbb{P}\left(\pi^0(\varepsilon x) > x\right) = \mathrm{o}\left(x^{1-\beta}\right)$$

holds. From this result, it follows that, given δ, one can choose sufficiently large β and sufficiently small ε so that

$$\mathrm{III} = \mathbb{P}\left(\pi^0(\varepsilon x) > \delta x\right) = \mathrm{o}(\mathbb{P}(B > x)).$$

Set $T_n = \pi^1(\varepsilon x) + \cdots + \pi^n(\varepsilon x)$. It remains to be shown that

$$\mathrm{IV} = \mathbb{P}\left(T_{\lfloor ax \rfloor} > (1-\delta)x\right) = \mathrm{o}(\mathbb{P}(B > x)).$$

Let $q > 0$. The last probability is smaller than

$$\lfloor ax \rfloor \mathbb{P}\left(\pi^1(\varepsilon x) > qx\right) + \mathbb{P}\left(\pi^1(\varepsilon x) \wedge qx + \cdots + \pi^{\lfloor ax \rfloor}(\varepsilon x) \wedge qx > (1-\delta)x\right).$$

For a given q, the first term is of $\mathrm{o}(\mathbb{P}(B > x))$ if ε is chosen suitably small (w.r.t. q). The second term is smaller than

$$\mathbb{P}\left(\pi^1 \wedge qx + \cdots + \pi^{\lfloor ax \rfloor} \wedge qx > (1-\delta)x\right),$$

with π^i now the ordinary busy period of the $M/G/\infty$ queue. Since I has finite p-th moment and is hence regularly varying with index p at infinity, the same holds for π^i, see e.g. Proposition 1 of Daley [8]. The proof can then be completed with the same kind of arguments as those used in the proof of Lemma 2. □

6. Reneging behavior of large customers

In this section, it is assumed that impatience is linear, i.e. $I = \theta B$ for $\theta > 1$. The probability of reneging of a customers when its service time is large is investigated. Let $L_{\theta,N}(x)$ denote the probability of reneging of a customer with service time equal to x. Since $I = \theta B$, we have

$$L_{\theta,N}(x) = \mathbb{P}\left(R(\theta x) < x\right). \tag{14}$$

As shown in the previous section, $R(x)/x \to \gamma_{N,\theta}$ for some constant $\gamma_{N,\theta}$.

The value of $\theta\gamma_{\theta,N}$ is an important parameter. If it is smaller than 1, then $L_{\theta,N}(x)$ converges to 1 as x gets large. This is undesirable, because of the overhead due to the reneging of large customers.

Thus, for a given value of θ, in order to prevent the reneging of large files one would like to choose N as large as possible so that $\theta\gamma_{\theta,N} > 1$. It is quite natural to define

$$N^* = \min\{N : \theta\gamma_{\theta,N} > 1\}.$$

Unfortunately, computing $\gamma_{\theta,N}$ explicitly turns out to be quite difficult. For an approximation procedure, see Boyer *et al.* [4]. Note also that $\gamma_{\theta,N}$ is lower-bounded by $\gamma_{\infty,N}$ given by Equation (9). This bound can be used to obtain a sufficient condition under which $\theta\gamma_{\theta,N} > 1$.

If the value of $\theta\gamma_{\theta,N}$ is only slightly larger than 1, it may happen that large customers still reneg with reasonable probability. This could be a reason to decrease N further. Notice that the impatience is completely removed when $N = N_* = \lfloor\theta\rfloor$.

The value $N = N_*$ may be too conservative, while the value N^* may lead to an undesirably high reneging rate. To obtain some insight in what happens between these two extreme cases, we study the asymptotic behavior of $L_{\theta,N}(x)$, for fixed θ and N. More generally, the asymptotic behavior of the probability $\mathbb{P}(R(x) < ax)$ for fixed a is investigated. Before doing so, some heuristic arguments, very similar to those in Section 3, are given.

As one can expect from the analysis carried out in earlier sections, the asymptotic behavior of this probability is determined by the average service rate in the PS queue with reneging and blocking, and k additional permanent customers. Denote this average service rate by γ_k (the parameters θ and N are omitted for simplicity). The event $\{R(x) < ax\}$ takes place if a certain number of other large customers are in the system simultaneously. In particular, we need to choose $k = k(a)$ so that

$$k(a) = \inf\{k : \gamma_{k+1} < a\}. \tag{15}$$

The main result of this section is the following theorem.

Theorem 5. *If $N < \infty$ and that the condition $\gamma_{\theta,N}^{k(a)} > a$ holds, then*

$$\mathbb{P}(R(x) < ax) = O\left(\mathbb{P}(B^r > x)^{k(a)}\right)$$

where $k(a)$ is defined by Equation (15)

This theorem is complementing Theorem 2 where N and I are infinite. For this queue, the number of customers in the system can be stochastically upper-bounded by the number of customers in an $M/G/\infty$ queue with service times θB. Note that, without loss of generality, it can be assumed that $\theta < N$. The structure of the proof is very similar, although the detailed arguments are different.

Notation: As before, $C(\varepsilon, x)$ denotes the amount of (non-tagged) customers in the system with service time larger than εx. The quantity $C(\varepsilon, x)$ counts both those customers in the system at time 0 with remaining service time larger than εx and the number of customers entering the system between time 0 and x with service time larger than the quantity εx.

As before, Q_ℓ denote the number of non-permanent customers in a queue with ℓ permanent customers; the notation $Q_{\ell,<\varepsilon x}$ is used to indicate that all service times are bounded by εx (i.e. distributed as $B \mid B < \varepsilon x$). For these variables, the parameters θ and N are omitted but implicit. Finally, $\tau_\ell(\varepsilon, x)$ is defined as the number of busy cycles of this queue between 0 and x. Note that a cycle ends if a non-permanent customer leaves the system which is then occupied with only the permanent customers.

To prove Theorem 5, a series of technical lemmas are derived. We first prove a technical lemma which describes the asymptotic behavior of $\tau_\ell(\varepsilon, x)$ when x goes to infinity.

Lemma 5. *Suppose $N < \infty$ or $\theta < \infty$. For any $b > 0$, and any $\beta > 0$ there exists an ε such that*

$$\mathbb{P}\left(\tau_\ell(\varepsilon, x) < bx\right) = o(x^{-\beta}).$$

Proof. Let $\pi_\ell(\varepsilon, x)$ denote the duration of a busy period of the queue with ℓ permanent customers and with reneging and blocking. Write

$$\mathbb{P}\left(\tau_\ell(\varepsilon, x) < bx\right) = \mathbb{P}\left(\pi_\ell^0(\varepsilon, x) + \sum_{i=1}^{\lfloor bx \rfloor} \pi_\ell^i(\varepsilon, x) > x\right),$$

where $\pi_\ell^i(\varepsilon, x)$, $i = 1, \dots, \lfloor bx \rfloor$ are independent copies of $\pi_\ell(\varepsilon, x)$ and $\pi_\ell^0(\varepsilon, x)$ is the residual busy period at time 0.

To upper-bound the above probability, one can proceed exactly as in Theorem 4. An upper-bound for $\mathbb{P}\left(\pi_\ell^0(\varepsilon, x) > x\right)$ and $\mathbb{P}\left(\pi_\ell(\varepsilon, x) > x\right)$ can be obtained by defining an appropriate $M/G/\infty$ queue: It is obvious that $\pi_\ell^0(\varepsilon, x)$ and $\pi_\ell(\varepsilon, x)$ are respectively upper-bounded by the residual and the ordinary busy period of an $M/G/\infty$ queue in which customers have service times distributed as $(I \mid I < \theta\varepsilon x)$. One can then proceed exactly as in the proof of Theorem 4. This completes the proof. □

The following lemma is an analogue of Lemma 2 for this situation.

Lemma 6. *If $\theta < \infty$ or $N < \infty$, then for each $\beta > 0$ there exists $\varepsilon > 0$ such that*

$$\mathbb{P}\left(R(x) < ax, C(\varepsilon, x) \leq k(a) - 1\right) = o(x^{-\beta}).$$

Proof. The amount of service received by a large customer when $C(\varepsilon, x) \leq k(a) - 1$ is at least the amount of service received by a large customer in a PS queue when there are $k(a)$ permanent customers. Thus, let us consider

$$R_k(x) = \int_0^x \frac{1}{k + Q_k(u)}\, du.$$

By using an ergodic theorem for the underlying Harris Markov chain, one gets that the quantity $R^k(x)/x$ converges a.s. to γ_k. Note that $\gamma_{\theta,N}^k$ is strictly decreasing with k. Note that $\gamma_{\theta,N}^N = 1/N$.

By a monotonicity argument (by assuming that all large customers are in the system permanently between time 0 and x), if $C(\varepsilon, k) \leq k(a)-1$, the quantity $R(x)$ can be lower-bounded by

$$R_{k(a),<\varepsilon x} = \int_0^x \frac{1}{k(a)+Q_{k(a),<\varepsilon x}(u)}\,du.$$

By using exactly the same arguments as in the proof of Theorem 4, we have

$$R_{k(a),<\varepsilon x} \geq \frac{1}{k(a)+1}\sum_{i=1}^{\tau_{k(a)}(\varepsilon,x)} E_i,$$

where E_i, $i=1,\ldots,\tau_{k(a)}$ are independent exponential random variables with parameter λ. We then deduce that

$$\begin{aligned}\mathbb{P}\left(R(x)<ax\right) &\leq \mathbb{P}\left(\frac{1}{k(a)+1}\sum_{i=1}^{\tau_{k(a)}(\varepsilon,x)} E_i < ax\right)\\ &\leq \mathbb{P}\left(\tau_{k(a)}(\varepsilon,x)<\lfloor bx\rfloor\right)+\mathbb{P}\left(\frac{1}{k(a)+1}\sum_{i=1}^{\lfloor bx\rfloor}E_i<ax\right)\end{aligned}$$

for some $b>0$. As in the proof of Theorem 4, the second term on the r.h.s. of the above inequality tends exponentially fast to 0 when $b>(k(a)+1)a\lambda$. From Lemma 6, for every $\beta>0$ the variable ε can be chosen so that $\mathbb{P}\left(\tau_{k(a)}(\varepsilon,x)<bx\right)$ is $o(x^{-\beta})$. As a consequence, the intermediate parameter b can be taken so that for every $\beta>0$ there exists an $\varepsilon>0$ so that

$$\mathbb{P}\left(R(x)<ax, C(\varepsilon,x)\leq k(a)-1\right)=o(x^{-\beta}).$$

This completes the proof. □

We finally proceed to the proof of Theorem 5.

Proof of Theorem 5. By using Lemma 6, it follows immediately that for $\beta>0$ and for suitably small ε,

$$\mathbb{P}\left(R(x)<ax\right)=\mathrm{o}(x^{-\beta})+\mathbb{P}\left(R(x)<ax, C(\varepsilon,x)\geq k(a)\right).$$

The last term is smaller than $\mathbb{P}\left(C(\varepsilon,x)\geq k(a)\right)$. If C_0 is the number of large customers present in the system at time 0, and $C_{(0,x]}$ is the number of arrivals with service time larger than εx (blocked and non-blocked), then $C(\varepsilon,x)\leq C_0+C_{(0,x]}$. Regardless of the particular model under consideration, the variable $C_{(0,x]}$ has a Poisson distribution with rate $\lambda x\mathbb{P}\left(B>\varepsilon x\right)$.

Next, consider C_0. The distribution of this quantity is not known. However, it is still possible to obtain a stochastic upper-bound for C_0: Consider an $M/G/\infty$ queue with service times θB. Let $\widetilde{Q}(u)$ be its associated queue length process. The PS queues can be constructed from $\widetilde{Q}(u)$ by deleting customers at appropriate places. Deleting occurs when customers are blocked or when customers leave because they have completed their service or they have reneged.

In the $M/G/\infty$ queue, customers are divided in two types according to the value of the initial service time, less or greater than εx. Denote respectively the number of these two types of customers by $(\widetilde{Q}^1(u))$ and $(\widetilde{Q}^2(u))$.

Then, it is obvious that $C_0 \leq \widetilde{Q}^1(0)$, and $\widetilde{Q}^1(0)$ has a Poisson distribution with parameter $\rho\mathbb{E}\left(B1(B > \varepsilon x)\right) = \rho x \mathbb{P}\left(B > \varepsilon x\right) + \rho\mathbb{P}\left(B^r > \varepsilon x\right)$. Hence,

$$\begin{aligned}\mathbb{P}\left(C(\varepsilon, x) \geq k(a)\right) &\leq \mathbb{P}\left(C_{[0,x]} \geq k(a)\right) + \sum_{k=0}^{k(a)-1} \mathbb{P}\left(C_{[0,x]} = k\right)\mathbb{P}\left(C_0 \geq k(a) - k\right) \\ &\leq \mathbb{P}\left(C_{[0,x]} \geq k(a)\right) + \sum_{k=0}^{k(a)-1} \mathbb{P}\left(C_{[0,x]} = k\right)\mathbb{P}\left(\widetilde{Q}^1(u) \geq k(a) - k\right)\end{aligned}$$

The proof can then be completed by using arguments similar to those used in the proof of Lemma 3. Details are omitted. □

References

1. Søren Asmussen, *Applied probability and queues*, John Wiley & Sons Ltd., Chichester, 1987.
2. Søren Asmussen, Claudia Klüppelberg, and Karl Sigman, *Sampling at subexponential times, with queueing applications*, Stochastic Process. Appl. **79** (1999), no. 2, 265–286.
3. T. Bonald and A. Proutière, *Insensitivity in processor-sharing networks*, Performance and Evaluation **49** (2002), 193–209.
4. Jacqueline Boyer, Fabrice Guillemin, Philippe Robert, and Bert Zwart, *Heavy tailed M/G/1-PS queues with impatience and admission control in packet networks*, Infocomm'2003 (San Francisco, USA), 2003.
5. D. Clark, S. Shenker, and L. Zhang, *Supporting real time applications in an integrated services packet network: architecture and mechanism*, Proc. Sigcomm'92 (Baltimore), 1992.
6. E. Coffman, A.A. Puhalskii, M.I. Reiman, and P.E. Wright, *Processor shared queues with reneging*, Performance Evaluation **19** (1994), no. 1, 25–46.
7. J.W. Cohen, *The multiple phase service netwrok with generalized processor sharing.*, Acta Informatica **12** (1979), 245–284.
8. D. J. Daley, *The busy period of the $M/GI/\infty$ queue*, Queueing Systems. Theory and Applications **38** (2001), no. 2, 195–204.
9. S. Foss and D. Korshunov, *Sampling at a random time with a heavy-tailed distribution*, Markov Process. Related Fields **6** (2000), no. 4, 543–568.
10. P. Jelenković and P. Momčilović, *Large deviation analysis of subexponential waiting time in an $M/G/1$ PS queue*, 2001, `http://comet.ctr.columbia.edu/~petar/ps.pdf`. Submitted.
11. P. Jelenković, P. Momčilović, and A.P. Zwart, *Reduced-load equivalence and subexponentiality.*, INRIA research report RR4444, 2001, Submitted.
12. F. P. Kelly, *Networks of queues*, Advances in Appl. Probability **8** (1976), no. 2, 416–432.
13. L. Kleinrock, *Queueing systems*, J. Wiley, New-York, 1976.
14. R. Núñez-Queija, *Processor sharing models for integrated services networks*, Ph.D. thesis, Eindhoven University of Technology, 2000.
15. S. Resnick and G. Samorodnitsky, *Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues*, Queueing Systems **33** (1999), 43–71.
16. J. L. van den Berg and O. J. Boxma, *The $M/G/1$ queue with processor sharing and its relation to a feedback queue*, Queueing Systems Theory Appl. **9** (1991), no. 4, 365–401.
17. A. P. Zwart, *Tail asymptotics for the busy period in the $GI/G/1$ queue*, Math. Oper. Res. **26** (2001), no. 3, 485–493.
18. A. P. Zwart and O. J. Boxma, *Sojourn time asymptotics in the $M/G/1$ processor sharing queue*, Queueing Systems Theory Appl. **35** (2000), no. 1-4, 141–166.

(Fabrice Guillemin) FRANCE TELECOM R&D, DAC/CPN, 22300 LANNION, FRANCE
E-mail address: Fabrice.Guillemin@rd.francetelecom.fr

(Philippe Robert) INRIA-ROCQUENCOURT, RAP PROJECT, DOMAINE DE VOLUCEAU, 78153 LE CHESNAY, FRANCE
E-mail address: Philippe.Robert@inria.fr

(Bert Zwart) EINDHOVEN UNIVERSITY OF TECHNOLOGY, DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE, HG 9.35 P.O. BOX 513 5600 MB EINDHOVEN THE NETHERLANDS
E-mail address: zwart@win.tue.nl