The impact of reneging in processor sharing queues

The Impact of Reneging in Processor Sharing Queues

H. Christian Gromoll *
Department of Mathematics
Stanford University
Stanford, CA
94305-2125, USA
gromoll@math.stanford.edu

Philippe Robert
INRIA-Rocquencourt
RAP project
Domaine de Voluceau
78153 Le Chesnay, France
Philippe.Robert@inria.fr

Bert Zwart†
Eindhoven University
Department of Mathematics and Computer Science
P.O. Box 513
5600 MB Eindhoven, the Netherlands
zwart@win.tue.nl

Richard Bakker
Eindhoven University
Department of Mathematics and Computer Science
P.O. Box 513
5600 MB Eindhoven, the Netherlands
R.F.Bakker@student.tue.nl

ABSTRACT

We investigate an overloaded processor sharing queue with renewal arrivals and generally distributed service times. Impatient customers may abandon the queue, or renege, before completing service. The random time representing a customer's patience has a general distribution and may be dependent on his initial service time requirement. We propose a scaling procedure that gives rise to a fluid model, with nontrivial yet tractable steady state behavior. This fluid model captures many essential features of the underlying stochastic model, and we use it to analyze the impact of impatience in processor sharing queues. We show that this impact can be substantial compared with FCFS, and we propose a simple admission control policy to overcome these negative impacts.

Categories and Subject Descriptors

C.4 [**Computer System Organization**]: Performance of Systems

General Terms

Algorithms,Performance

*Research supported in part by an NSF Mathematical Sciences Postdoctoral Research Fellowship, a European Union Marie Curie Postdoctoral Research Fellowship, and EURANDOM
†Research supported by an NWO-VENI grant

SIGMetrics/Performance'06, June 26–30, 2006, Saint Malo, France.

Keywords

Processor Sharing. Queues in Overload. Queues with Impatience. Measure Valued Process. Fluid Limits. Delay-Differential Equation. User behavior. Admission control.

1. INTRODUCTION

Over the past few years, flow level models have become an important tool in the analysis of bandwidth sharing networks carrying elastic traffic; see for example Kelly [15], Massoulié and Roberts [19] and Mo & Walrand [20]. In such models, elastic flows share the system resources fairly, which coincides naturally with the Processor Sharing (PS) discipline when studying a single link carrying identical flows. The single server PS queue with Poisson arrivals and generally distributed service times has very tractable steady state behavior: the distribution of the number of flows in the system is insensitive to the service time distribution (apart from its mean). This has led to considerable renewed interest in the analysis of PS queues.

Recently, Bonald and Proutière [3] have shown that this insensitivity property also holds in several bandwidth sharing networks, and that, in other cases, it is possible to obtain insensitive upper and lower bounds to the steady state distribution; see [4]. Kelly and Williams [16] take a different approach to evaluating the performance of bandwidth sharing networks. They propose to approximate a bandwidth sharing network with a more tractable fluid model. The same approach is taken in recent papers by Kang *et al.* [14], and Key *et al.* [17].

In the present paper, we take a similar approach and use a fluid model to analyze the performance of a $GI/GI/1$ PS queue with *impatient customers*, abbreviated PSI queue. In a PSI queue, a customer departs the queue either when it has completed service, or when its sojourn time has exceeded a certain *reneging time*. In the latter case, the customer is lost due to impatience. Such behavior may correspond to actual impatient customers, or by a time out caused by TCP or higher layer protocols.

Reneging under PS has, a priori, a larger impact than under FCFS, since in the latter case, customers typically abandon the queue before beginning service. This is not the case under PS, where jobs that renege have already received some amount of service. Thus, impatience may create significant overhead for a PS system. This is especially true if the system is overloaded, a condition that can occur (at least temporarily) in practice. The main goal of this paper is to develop tools to investigate the impact of this type of user behavior.

There is a large literature on queueing models with impatient customers under the FCFS discipline. An early paper is Barrer [1], where a queue with impatient customers arises in a military application. For a survey, see Stanford [23]. (See also Stanford [22] and Boots and Tijms [6].) This literature focuses on exact performance analysis of the system involved. A diffusion limit for single channel queues has been obtained recently by Ward and Glynn [25]. There are also various studies of multiserver queues with reneging, motivated by call center applications; see the survey by Gans *et al.* [9] and references therein.

The literature on PS queues with impatience shows a different picture: there are few results available. An exception is the case with exponentially distributed service times, and independent exponential reneging times; see Coffman *et al.* [7]. Guillemin *et al.* [12] consider PS queues with impatient customers and heavy tailed service times, and obtain some results on the reneging behavior of large jobs by analyzing the tail behavior of the steady state sojourn time distribution. In a series of papers, Doytchinov and coauthors [8, 18] have investigated heavy traffic limits of queues with impatience, under a variation of the Earliest Deadline First policy. In this case, impatient customers do not abandon the queue and the quantity of interest is their lateness when completing service.

Using some approximations, Bonald and Roberts [5] analyze the steady state of a system with general service times and some dependence between service times and reneging times. It is shown by simulation in [5] that, if customers are relatively patient (compared to the speed of the server), the service rate becomes approximately constant in steady state, which facilitates their approximations. We complement these simulation results by proving that their approximations are exact on fluid scale. In addition, we also consider the time dependent behavior of the system and we do not make any a priori assumptions about the distribution of the service times or the reneging times.

In a related paper, Yang and De Veciana [26] show by simulations that user impatience can have a substantial impact on the performance of the system, especially when the system is overloaded. They consider a more complex model of customer impatience than we do here. Section 6 indicates how our model could be extended towards their level of generality.

We now discuss the approach taken in the present paper. In Sections 2 and 3 we introduce the PSI queue, and introduce a fluid scaling by speeding up the arrival and service rates by a factor r. We show that our rescaled PSI queue converges in distribution to a fluid limit. The fluid limit is described by the solution of a functional fixed point equation, which can be seen as a time changed delay-differential equation.

This fluid limit exhibits non-trivial but tractable steady state behavior if the system is overloaded. The steady state behavior is completely characterized by a simple fixed point equation. This equation provides considerable information about the performance of the system, as is illustrated by several examples. For example, we prove that more variability in the service times and/or reneging times produces better system performance, which is in accordance with results in [5]. In addition, we also investigate the time dependent behavior of the system, by numerically solving the fluid model equation. It seems that steady state is reached fairly quickly when either the service time or reneging time distribution is light tailed. If both are heavy tailed, then convergence is slow, but impatience is still shown to have a significant effect on system performance.

To reduce the impact of reneging, one may proceed in various ways. One way, proposed by Yang and De Veciana [26], is to use size based scheduling disciplines instead of PS. In the present work, we follow the approach suggested in [5] and focus on a simple admission control policy: we assume that arrivals are blocked when the total number of jobs in the system exceeds a certain threshold. Using a heuristic interpretation obtained in Section 4, we show how one can evaluate the performance of this system. It is shown that admission control always leads to increased "goodput" and often (but not always) to an increased number of successfully transmitted jobs.

On a technical level, the model considered here poses some challenges. In contrast to Doytchinov *et al.* [8], Gromoll *et al.* [11] and Puha *et al.* [21], the service discipline is not work conserving and, for this reason, analysis of the fluid model is more intricate. This is an important difference from earlier work on standard PS queues, where the fact that the workload process coincides with that of FCFS queues plays an important role. A different approach to prove existence, uniqueness, and convergence to steady state of fluid model solutions is used. By iteratively defining *minimal* and *maximal* solutions, and by using monotonicity arguments, we are able to investigate the properties of the fluid limits under quite general assumptions. To show that the fluid model is indeed a fluid limit of the original PSI queue, we use, among other techniques, results from empirical process theory.

The paper is organized as follows. In Section 2 we introduce the PSI queue. In Section 3, we describe the fluid scaling and present our main convergence results, which reduce the PSI queue to a more tractable fluid model equation. We show that under weak assumptions, this equation has a unique solution and has a nontrivial limiting behavior. The full technical details concerning the convergence to the fluid limit can be found in Gromoll *et al.* [10]. Readers may skip the technical details in Sections 2 and 3 and move directly to Section 4, where we apply the convergence results of Section 3 to analyze the performance of the PSI queue. Section 5 is concerned with admission control. Extensions such as reattempts and more complex user behavior are discussed in Section 6. Section 7 concludes.

2. MODEL DESCRIPTION

We consider a processor sharing server working at unit rate with an infinite capacity buffer. Let $(E(\cdot), (B_i, D_i))$ be a collection of random elements describing respectively the arrival process, the service times, and the corresponding reneging times. The expectation with respect to these variables is denoted $\mathbf{E}$.

For $t \geq 0$, $E(t)$ is the number of arrivals up to time t. It is assumed that $E(\cdot)$ is a (possibly delayed) renewal process with intensity $\lambda \in (0, \infty)$. U_i is the arrival time of the ith job, $i \geq 1$. Jobs already in the buffer at time 0 will be called *initial jobs*.

The sequence (B_i, D_i) is assumed to be independent, identically distributed (i.i.d.) with common joint distribution ϑ on $[0, \infty] \times [0, \infty]$. For $i \geq 1$, B_i is the amount of processing time that job i requires from the server. The random variable D_i determines the deadline of job i: since it arrives at time U_i, it must complete service before its deadline at time $U_i + D_i$. The variable D_i is called the reneging time of job i. Note that, for $i \geq 1$, the variables B_i and D_i are allowed to be dependent. Note also that for each i, we allow either B_i or D_i to equal infinity. This allows us to incorporate the standard PS queue and the $GI/GI/\infty$ queue as special cases, as well as other useful examples; see Section 4.3.

2.1 A measure valued process

At time $t \geq 0$, a job in the queue has two characteristics: its *residual service time* b, representing the remaining amount of processing time it requires to complete service, and its current *lead time* d, representing the remaining time until its deadline expires. This will be represented as the point (b, d) of $[0, \infty] \times [0, \infty]$. Due to the processor sharing discipline, the first coordinate b decreases at rate $1/Z(t)$ if $Z(t)$ is the number of buffered jobs at time t. The second coordinate d decreases at rate 1; see Figure 1. The system can be described as a distribution of points on $[0, \infty] \times [0, \infty]$ moving toward the axes. When a point of this distribution hits one of the axes, it disappears: if it hits the vertical axis, its residual service time equals zero and the job departs the queue due to service completion; if the point hits the horizontal axis, the current lead time of the job equals zero and the job is lost due to reneging. Since the reneging time D_i of job i is equal to its current lead time at time $t = 0$, the reneging time is also referred to as the *initial* lead time.

To keep track of the evolution of the system, on must know the location of all points. Since there is no upper bound on the total number of jobs, the state space is infinite dimensional. A convenient way to deal with this is by using a *measure valued process*. Informally, we have a process $\mathcal{Z}(t)$, $t \geq 0$, such that $\mathcal{Z}(t)(F)$ counts the number of currently buffered jobs with residual service time and current lead time in the set F. If $F = [0, \infty] \times [0, \infty]$, then we get the total number of jobs in the system. In the rest of this section, a more precise description of the measure valued process and notation is introduced.

The *initial condition* specifies $Z(0)$, the number of initial jobs present in the buffer at time zero, as well as the service time requirements and initial lead times of these initial jobs. Assume that $Z(0)$ is a nonnegative, integer valued random variable. The service times and initial lead times for initial jobs are the first $Z(0)$ elements of an i.i.d. sequence $(\tilde{B}_j, \tilde{D}_j)$. It is assumed that $Z(0)$ and $(\tilde{B}_j, \tilde{D}_j)$ are independent of $E(\cdot)$ and the sequence (B_i, D_i). A convenient way to express the initial condition is to define an initial random measure $\mathcal{Z}(0)$ on $[0, \infty] \times [0, \infty]$,

$$\mathcal{Z}(0) = \sum_{j=1}^{Z(0)} \delta_{(\tilde{B}_j, \tilde{D}_j)},$$

where δ_X is the Dirac mass at X. Henceforth, $\mathcal{Z}(0)$ will be used as the description of the initial condition. It is assumed that

$$\mathbf{E}[\langle 1, \mathcal{Z}(0) \rangle] = \mathbf{E}[Z(0)] < \infty, \tag{2.1}$$

where $\langle f, \mu \rangle$ denotes the integration of f with respect to the measure μ.

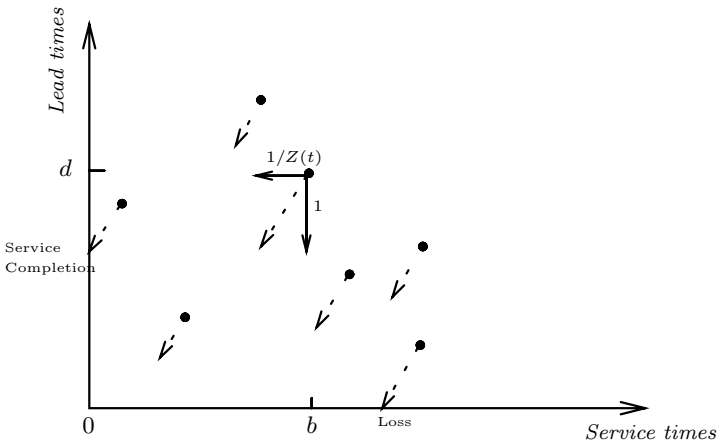

Figure 1: Dynamics of the points of the measure valued process $\mathcal{Z}(\cdot)$

Cumulative service per job

For each $t \geq 0$, let $S(t)$ denote the *cumulative service per job* provided by the server up to time t. Thus, if job i (arrived at time U_i) is still in the queue at time t, the cumulative amount of processing it receives by time t equals $S(t) - S(U_i)$. With this definition, job i leaves the queue at time

$$\inf\{s \geq U_i : S(s) - S(U_i) \geq B_i \text{ or } s - U_i \geq D_i\}.$$

If the queue is not empty during $[0, t]$ and if the process $Z(\cdot)$ tracks the number of customers in the queue, the quantity $S(t)$ can be expressed as

$$S(t) = \int_0^t \frac{1}{Z(s)}\, ds.$$

Similarly, the cumulative amount of processing time received by an initial job $j \leq Z(0)$, still buffered at time t, is $S(t)$.

The *residual service time* $B_i(t)$ of job $i \leq E(t)$ at time t, resp. of initial job $j \leq Z(0)$ is defined as, respectively,

$$B_i(t) = (B_i - (S(t) - S(U_i)))^+, \quad \tilde{B}_j(t) = (\tilde{B}_j - S(t))^+.$$

The *lead time* $D_i(t)$ of job $i \leq E(t)$ at time t, resp. of initial job $j \leq Z(0)$ is given by, respectively,

$$D_i(t) = U_i + D_i - t, \qquad \tilde{D}_j(t) = \tilde{D}_j - t. \tag{2.2}$$

The state descriptor

We denote the space of finite, non-negative Borel measures on the quadrant $\bar{\mathbb{R}}^2_+ = [0, \infty] \times [0, \infty]$ by M and for $X \in \bar{\mathbb{R}}^2_+$, the measure $\delta_X \in M$ is the Dirac mass at X with the convention that $\delta_X \equiv 0$ when one of the coordinates of x is zero (so that jobs that have departed the queue are not included in the state description).

The PSI queue at time t is represented as a random element of M as follows: at time t, $\mathcal{Z}(t)$ has a unit of mass located at $(\tilde{B}_j(t), \tilde{D}_j(t)) \in \bar{\mathbb{R}}^2_+$ for each initial job $j \leq$

$Z(0)$ still in the buffer at time t, and a unit of mass at $(B_i(t), D_i(t)) \in \bar{\mathbb{R}}_+^2$ for each job $i \le E(t)$ still in the buffer at time t. Therefore, the random measure $\mathcal{Z}(t)$ can be expressed as

$$\mathcal{Z}(t) = \sum_{j=1}^{Z(0)} \delta_{(\tilde{B}_j(t), \tilde{D}_j(t))} + \sum_{i=1}^{E(t)} \delta_{(B_i(t), D_i(t))}. \tag{2.3}$$

Since $Z(t)$ denotes the number of jobs in the buffer at time $t \ge 0$, $Z(t) = \langle 1, \mathcal{Z}(t)\rangle$. We define $\varphi(x) = 1/x$ for $x \in (0, \infty)$ and $\varphi(0) = 0$. Due to the processor sharing discipline, a customer present in the queue during the time interval $[t, t+dt]$ receives the amount of service $\varphi(\langle 1, \mathcal{Z}(t)\rangle)dt$. The cumulative service $S(t)$ per job up to time t can therefore be written as

$$S(t) = \int_0^t \varphi(\langle 1, \mathcal{Z}(s)\rangle)\, ds. \tag{2.4}$$

Let f be a non-negative Borel measurable function on $\mathbb{R}^2$ whose support is in $(0, \infty] \times (0, \infty]$. Since

$$\langle f, \mathcal{Z}(t)\rangle = \int_{\mathbb{R}^2} f(u, v)\, \mathcal{Z}(t)(du, dv),$$

it is easily seen that the dynamics of the points of $\mathcal{Z}(\cdot)$ are given by

$$\langle f, \mathcal{Z}(t+dt)\rangle = \int f(B_{E(t)+1}, D_{E(t)+1})\, dE(t) + \int f(u - (S(t+dt) - S(t)), v - dt)\, \mathcal{Z}(t)(du, dv). \tag{2.5}$$

For a family of conveniently chosen functions f, this evolution equation plays a crucial role in determining fluid limits for the model.

3. CONVERGENCE TO A FLUID MODEL

We now scale our PSI queue with a scaling factor $r \in \mathcal{R}$, with $\mathcal{R}$ some sequence tending to infinity. To obtain nontrivial scaling limits in which the effects of reneging and successful service completions are both visible, we replace the lead times (D_i) by (rD_i). Informally, we let customers become relatively patient with respect to the service rate. In addition, we also speed up time by a factor r. An alternative way of looking at this scaling is to leave the time scale and reneging times unchanged and to speed up the arrival and service rates by a factor r. This procedure would lead to exactly the same fluid limit as described below.

For $r > 0$, the fluid scaled state descriptor is defined, for $t \ge 0$, as the random measure $\bar{\mathcal{Z}}^r(t) \in \mathbf{M}$ such that

$$\bar{\mathcal{Z}}^r(t)(F \times G) = \frac{1}{r}\mathcal{Z}(rt)(F \times rG),$$

for all Borel sets $F, G \subset \bar{\mathbb{R}}_+$. Note that this definition scales the lead times by a factor r^{-1} as well.

The analysis of the renormalized processes $\bar{\mathcal{Z}}^r(\cdot)$, $r > 0$ involves fluid scaled versions of many of the processes introduced so far. For all $r \in \mathcal{R}$, $t \ge 0$, and $i = 1, \ldots, E^r(rt)$, define

$$\bar{E}^r(t) = \frac{1}{r}E(rt), \quad \bar{S}^r(t) = S^r(rt),$$
$$\bar{Z}^r(t) = \frac{1}{r}Z(rt), \quad \bar{B}_i^r(t) = B_i(rt),$$
$$\bar{D}_i^r(t) = \frac{1}{r}D_i(rt).$$

The fluid scaled process $\bar{S}^r(\cdot)$ plays an important role, and it will be convenient to have notation for its increments. Define

$$\bar{S}^r(s, t) = \bar{S}^r(t) - \bar{S}^r(s).$$

Assumptions

We assume that ϑ is a probability measure on $\bar{\mathbb{R}}_+^2$ such that

$$\vartheta\left(\{0\} \times \bar{\mathbb{R}}^+\right) = \vartheta\left(\bar{\mathbb{R}}^+ \times \{0\}\right) = \vartheta(\{\infty\}, \{\infty\}) = 0. \tag{3.1}$$

We further assume that there exists a deterministic measure ζ_0 on the quadrant $\bar{\mathbb{R}}_+^2$ with $\langle 1, \zeta_0\rangle = z_0 < \infty$, such that

$$\zeta_0\left(\{0\} \times \bar{\mathbb{R}}^+\right) = \zeta_0\left(\bar{\mathbb{R}}^+ \times \{0\}\right) = \zeta_0(\{\infty\}, \{\infty\}) = 0,$$

and such that, as $r \to \infty$,

$$\bar{\mathcal{Z}}^r(0) \to \zeta_0. \tag{3.2}$$

in the topology of weak convergence of measures.

3.1 Main results

In order to state our main results, we introduce some additional notation. Let (B, D) be a generic random element of $\bar{\mathbb{R}}_+^2$ with distribution ϑ, and let (B_0, D_0) be a random element of $\bar{\mathbb{R}}_+^2$ with distribution ζ_0/z_0. Let $z(t)$, $t \ge 0$, be a solution of the equation

$$z(t) = z_0 \mathbf{P}\left(B_0 > S(0,t), D_0 > t\right) + \lambda \int_0^t \mathbf{P}\left(B > S(s,t), D > t - s\right) \mathrm{d}s,\ t < t^*, \tag{3.3}$$

where $S(u, v) = \int_u^v (1/z(w))\mathrm{d}w$ and $t^* = \inf\{t > 0 : z(t) = 0\}$. We define $z(t) = 0$ for $t > t^*$.

For given $x, y \ge 0$ and a given process $z(t)$, we define

$$\zeta(t)(x, y) = \zeta(t)([x, \infty] \times [y, \infty])$$

as follows:

$$\zeta(t)(x,y) = z_0 \mathbf{P}\left(B_0 > x + S(0,t), D_0 > y + t\right) + \lambda \int_0^t \mathbf{P}\left(B > x + S(t-s,t), D > y + s\right) \mathrm{d}s. \tag{3.4}$$

Note that $z(t) = \zeta(t)(0, 0)$. This characterizes a measure valued function $\zeta(\cdot)$.

The function $z(\cdot)$ can be viewed as an approximation of the number of customers in the PSI queue on fluid scale (that is, of $\bar{Z}^r(\cdot)$). Analogously, $\zeta(\cdot)$ can be seen as an approximation of the measure valued process $\bar{\mathcal{Z}}^r(\cdot)$.

We are now ready to state our main results, it corresponds to Theorem 3.10 of Gromoll *et al.* [10]. The proof of the convergence of the renormalized process to a solution of Equation (3.4) is quite involved, mainly because of the complicated state space. We give a sketch of the proofs in the simple case of Poisson arrivals. The crucial starting point is the equation of evolution (2.5) which is rewritten by using some martingales associated to Poisson process. It turns

out that, with the renormalization, these martingales converge to 0 as r goes to infinity. In the paper, since general arrivals are considered, the proof of this result is more sophisticated. The second (intricate) step is to show that the sequence of renormalized measure valued processes is tight for the convergence in distribution (section 5 of [10]). The last step is a continuity result so that one can take the limit in Equation (2.5) written for the renormalized process (section 6 of [10]) to get Equation (3.4) , uniqueness results for the limiting equation are also required at this stage.

The first theorem deals with the behavior at infinity of solutions to the Equation (3.3).

THEOREM 3.1. *Suppose that*

$$\lambda \mathbf{E}\left[B1_{\{D=\infty\}}\right] < 1 \text{ and } \mathbf{E}\left[\min\{B, D\}\right] < \infty.$$

Any solution $z(\cdot)$ of (3.3) converges to 0 if $\lambda\mathbf{E}[B] < 1$, and to the unique positive solution z of the fixed point equation $z = \lambda\mathbf{E}[\min\{D, zB\}]$, if $\lambda\mathbf{E}[B] > 1$. In the latter case, any solution to (3.4) converges to the measure ζ_∞ given by

$$\begin{aligned}\zeta_\infty(x, y) &= \lambda \int_0^\infty \mathbf{P}\left(B > x + t/z, D > t + y\right) \mathrm{d}t \\ &= \mathbf{E}\left[\min\{z(B - x)^+, (D - y)^+\}\right].\end{aligned}$$

The simple fixed point equation $z = \lambda\mathbf{E}[\min\{D, zB\}]$ is quite tractable, and various properties of its solution are analyzed in the next section. The second main result of the present section deals with uniqueness of solutions to (3.4) (and consequently, uniqueness of solutions to (3.3)).

THEOREM 3.2. *Let $y' > y \geq 0$ and suppose there exists a constant L such that*

$$\zeta_0([y, y'] \times F) \leq L|y' - y| \tag{3.5}$$

for any Borel set F. Then (3.3) and (3.4) have a unique solution.

The justification of the function $\zeta(\cdot)$ as a fluid approximation of the original fluid scaled PSI queue described by $\bar{\mathcal{Z}}^r(\cdot)$ is provided by the following result.

THEOREM 3.3. *Suppose Relations* (3.1) *and* (3.2) *hold.*

(i) The sequence of fluid scaled processes $\{\bar{\mathcal{Z}}^r(\cdot)\}$ is tight as $r \to \infty$.

(ii) Every limit point satisfies equation (3.4).

(iii) If in addition (3.5) holds, then $\bar{\mathcal{Z}}^r(\cdot)$ converges in distribution to $\zeta(\cdot)$ as $r \to \infty$.

The above theorems reduce an intricate measure valued process to a tractable fluid model. In the next section, we investigate this fluid model to analyze the impact of reneging on system performance.

4. PERFORMANCE ANALYSIS

In this section, we investigate the performance of the PSI queue by using the fluid model introduced in the previous section. A main feature of the fluid model is that, if the system is in overload, the fluid model exhibits a nontrivial steady state behavior. In particular the number of users $z(t)$ converges to the unique positive solution of the following simple fixed point equation.

$$z = \lambda\mathbf{E}\left[\min\{zB, D\}\right]. \tag{4.1}$$

The main goal of this section is to investigate this equation, and investigate its validity by doing transient analysis. We treat a number of examples which allow for explicit computations, and also obtain a number of stochastic ordering results. We investigate the time dependent behavior of $z(t)$ using both analytic and numerical methods. An equivalent version of (4.1) has been proposed as a direct approximation in [5]. In that paper, the assumptions made on B and D imply the existence of a maximum job size b^* such that customers renege if and only if $B \geq b^*$. In the present paper we do not make such an assumption.

Before we analyze Equation (4.1), we first give a heuristic interpretation. Let Z^r denote the steady state number of customers in the rth system. Furthermore, let $V^r(B)$ be the sojourn time of a customer if the customer never reneges. Then the actual sojourn time is given by $\min\{V^r(B), Dr\}$, and from Little's law we get

$$\mathbf{E}\left[Z^r\right] = \lambda\mathbf{E}\left[\min\{V^r(B), Dr\}\right]. \tag{4.2}$$

Divide both sides of (4.2) by r. Since we observe the system in steady state at time 0, the number of customers hardly changes and by the so called "snapshot principle" we conclude that $V^r = Z^r B + \mathrm{o}(r)$. Noting that $Z^r/r \to z$ then gives (4.1) after dividing both sides of (4.2) by r and letting $r \to \infty$.

Apart from the mean queue length z, we are also interested in the long term fraction P_s of customers that leave the system successfully, and the "goodput" and "badput", i.e. the fractions of work by the server dedicated to successful and unsuccessful transfers. It is clear that

$$P_s = \mathbf{P}\left(D > zB\right).$$

The "goodput" T_s is given by $T_s = \lambda\mathbf{E}[B; zB < D]$. Another performance measure we are interested in is the reneging rate $d(t)$ at time t and the stationary reneging rate d. These are given by

$$\begin{aligned}d(t) &= z_0 f_{D_0}(t)\mathbf{P}\left(B_0 > S(0, t) \mid D_0 = t\right) \\ &\quad + \lambda \int_0^t \mathbf{P}\left(B > S(t - s, t) \mid D = s\right) d\mathbf{P}\left(D \leq s\right), \\ d &= \lambda \int_0^\infty \mathbf{P}\left(zB > s \mid D = s\right) d\mathbf{P}\left(D \leq s\right) \\ &= \lambda(1 - P_s)\end{aligned}$$

with $f_{D_0}(t)$ the marginal density of D_0.

The following remarkable property, which simply follows from the fixed-point equation (4.1), shows that the performance of the system does not depend on the average of D.

PROPERTY 4.1. *Consider two systems numbered by* 1 *and* 2 *such that $(B_2, D_2) \equiv (B_1, aD_1)$ for some $a > 0$, and such that $\lambda_1 = \lambda_2$. Then (with obvious notation) we have*

$$\begin{aligned}z_2 &= az_1, \\ P_{s,2} &= P_{s,1}.\end{aligned}$$

This property, which may seem surprising at first sight, can be explained as follows: Suppose that the system is in equilibrium at time 0 and suppose further that the arrival process at time 0 changes in such a way that the customers

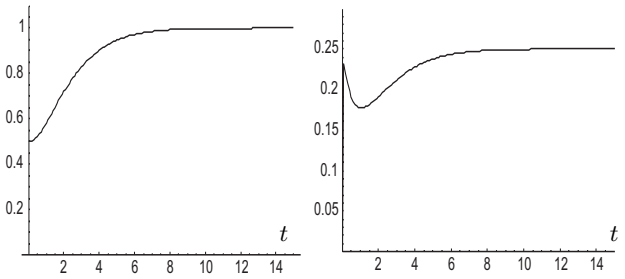

Figure 2: Illustration of Property 4.1: $z(t)$ (left) and $d(t)$ (right)

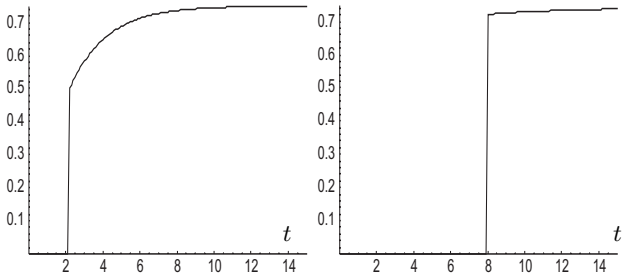

Figure 3: Reneging rate for $D = B$, exponential(left) and Pareto(right).

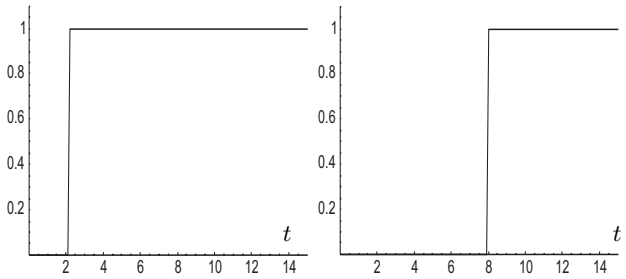

Figure 4: Fraction of departures due to reneging for $D = B$, exponential(left) and Pareto(right).

arriving after time 0 are twice as patient as the customers arriving before time 0. This makes the number of customers in the system grow larger until a new equilibrium has been reached. This new equilibrium occurs when there are twice as many customers in the system compared with the old situation. In the new equilibrium, customers are twice as patient, but the service rate is halved. Thus, the fraction of impatient customers stays the same. An illustration of this is given in Figure 2, where we consider a PSI queue where B_0 and B are exponentially distributed with rate $\mu = 1$, D_0 is exponentially distributed with rate $\nu_0 = 2$, and D is exponentially distributed with rate $\nu = 1$. z_0 is set equal to $(\lambda - \mu)/\nu_0$ (the equilibrium if D would be replaced by D_0). We take $\lambda = 2$. The figure shows that the system transfers from the old to the new equilibrium in a period which is 8 times the average service time.

In the remaining part of this section, we examine several different scenarios. In Section 4.1, we assume a strong form of dependence between B and D. In Section 4.2, we assume that B and D are independent of one another. Section 4.3 is an illustration of the fact that the model we consider is general enough to incorporate TCP friendly traffic. All these sections focus on the overloaded case $\rho > 1$. In fact, we take $\rho = 1.5$ and $z_0 = 0$ in all remaining numerical examples; other values of ρ and z_0 give similar insights.

In our analysis, we do not restrict to steady-state analysis by means of the fixed point equation (4.1), but also investigate the whole process $z(t)$ and other performance measures as mentioned above. In general, it is not possible to obtain a solution of $z(t)$, and therefore we compute $z(t)$ numerically using Picard iteration. An exception is the case where D has an exponential distribution, independent of B and $z_0 = 0$. For this case, we found an exact expression that is remarkably simple.

4.1 Completely dependent lead times

Consider first the case where $D = \Theta B$, with $\Theta > 0$ (independent of B) reflecting the average service rate expected by a customer. In this case, the performance measures can be determined from the equations (recall that $\rho = \lambda \mathbf{E}[B] > 1$)

$$z = \rho \mathbf{E}[\min\{\Theta, z\}],$$
$$P_s = \mathbf{P}(\Theta > z).$$

Some specific examples:

Θ *single-valued.*

If we assume that $\Theta = \theta$, then $z = \rho \min\{\theta, z\}$, which implies that $z = \rho\theta$ since $\rho > 1$. From this, it follows that all customers leave the system impatiently: $P_s = \mathbf{P}(\theta > \rho\theta) = 0$. Observe that when a customer leaves the system, a fraction $1/\rho$ of his service time has been processed.

Of course, this behavior is very undesirable. One may wonder if this only happens in steady-state. The next figure shows the reneging rate in the two cases where $z_0 = 0$, $\lambda = 3/4$, $B = D$, and B is either exponentially distributed with rate $\mu = 1/2$ or has a Pareto distribution, i.e. $\mathbf{P}(B > x) = (a/(a+x))^b$ with $b = 1.5, a = 1$ (to make the mean $a/(b-1) = 2 = 1/\mu$.

The reneging rate $d(t)$ in this case simplifies to

$$d(t) = \lambda \int_0^t f_B(s) I(S(t-s,t) \leq s) ds,$$

with f_B the density of B.

In both cases, the limiting reneging rate is equal to λ, the arrival rate. Figure 3 shows that the convergence strongly depends on the growth rate of $z(t)$ as long as $z(t) < 1$ (in which case there is no reneging), and the tail of the service time distribution. If service times are exponentially distributed, the growth rate is $\lambda - \mu = 1/2$. For Pareto service times, the growth rate is the solution of the equation $\rho \mathbf{E}\left[e^{sB^*}\right] = 1$, cf. [13], and turns out to be much smaller. When $z(t)$ exceeds 1, the reneging rate shows a sudden increase, and then gradually converges to its limiting value, the speed of convergence depending on the tail of the service time distribution.

If we look at the fraction of customers that leave due to reneging (i.e. the reneging rate divided by the reneging rate plus the service completion rate), then we see an extremely sharp transition from 0 to 1 for both exponentially and Pareto distributed service times, as illustrated by Figure 4. From this we conclude that reneging behavior has a significant impact on finite time scale, irrespective of the tail of the service time distribution.

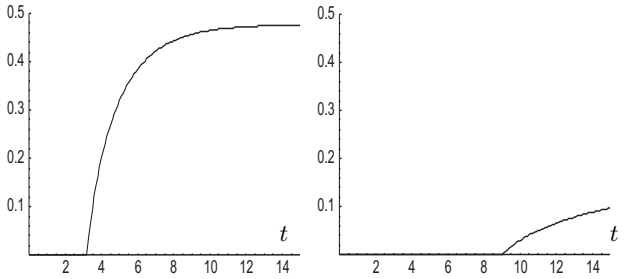

Figure 5: Reneging rate for $D = B + 1$, exponential(left) and Pareto(right).

Θ *two-valued.* From the previous example, it is clear that the system can only get some work done if some customers are more patient than others. In this example we assume that Θ equals θ_1 with probability p and θ_2 with probability $1-p$. Take $\theta_2 > \theta_1$. Equation (4.1) now simplifies to

$$z = \rho p \min\{z, \theta_1\} + \rho(1-p)\min\{z, \theta_2\}.$$

From this equation and the properties $\theta_2 > \theta_1, \rho > 1$ it follows that $z > \theta_1$. Furthermore, $z > \theta_2$ holds if and only if the equation $z = \rho p\theta_1 + \rho(1-p)z$ has a non-negative solution, which is the case if and only if $\rho(1-p) < 1$ (i.e. when the most patient customers cannot saturate the system alone). In this case we have

$$z = \frac{\rho p \theta_1}{1 - \rho(1-p)} < \theta_2.$$

If the last inequality is not valid or if $\rho(1-p) \geq 1$ we must have $z \geq \theta_2$ which implies $z = \rho p\theta_1 + \rho(1-p)\theta_2$. From the above we can conclude that $P_s = 0$ iff $(1-\rho(1-p))\theta_2 < \rho p\theta 1$. If the reverse inequality holds then all customers of type 2 are being served successfully, i.e. $P_s = (1-p)$.

Θ *exponentially distributed.* Assume w.l.o.g. that the mean of Θ equals 1. In this case z can be determined from the equation $z = \rho(1 - \mathrm{e}^{-z})$ and $P_s = \mathrm{e}^{-z} = 1 - z/\rho$.

Stochastic comparison

Since P_s does not depend on the mean of Θ, and since the worst-case property of the case of constant Θ, it seems natural to conjecture that the system performance is positively related to the variability of Θ. Thus it seems worthwhile to look for ordering relations for P_s if Θ_1 and Θ_2 are ordered in the convex ordering $\Theta_1 \geq_{cvx} \Theta_2$, i.e. $\mathbf{E}[f(\Theta_1)] \geq \mathbf{E}[f(\Theta_2)]$. This is well-known to be equivalent to $\mathbf{E}[\min\{x, \Theta_1\}] \leq \mathbf{E}[\min\{x, \Theta_2\}]$ for all $x \geq 0$.

Combining this with our fixed point equation for z, we have shown the following:

PROPOSITION 4.2. *if $\Theta_1 \geq_{cvx} \Theta_2$, then $z_2 \geq z_1$ i.e. less variability in reneging behavior implies a lower service rate.*

To prove that the loss rates are also ordered, i.e. that also $\mathbf{P}(\Theta_1 > z_1) \geq \mathbf{P}(\Theta_2 > z_2)$ seems hard without imposing further assumptions.

Grace period

In Yang and De Veciana [26] it is argued that customers have a certain initial "grace period", in which they will not leave the system due to impatience. This gives rise to the form $D = \Theta B + \Theta_1$, with Θ_1, Θ and B all independent. In Figure 5 we show the reneging rates for the case $D = B + 1$, with B exponentially distributed and Pareto distributed in the same way as before and $\lambda = 3/4$. Again the difference in convergence behavior is clear. The limiting values of $z(t)$ are 2.10303 in the exponential case and 1.92585 in the Pareto case, which gives rise to limiting reneging rates of $\lambda\mathbf{P}(B > 1/(z-1))$ which equals 0.4766 in the exponential case and 0.13874 in the Pareto case; a striking difference. This difference is in accordance with results reported in [5].

4.2 Independent lead times

As a second example, we now assume that D and B are independent. In this case we can write (4.1) as

$$\lambda \int_0^\infty \mathbf{P}(B > u)\,\mathbf{P}(D > zu)\,\mathrm{d}u = 1.$$

which, in case $\mathbf{E}[B] < \infty$, which we assume throughout this subsection, this is equivalent to $\mathbf{P}(D > zB^*) = 1/\rho$, with B^* a random variable with density $\mathbf{P}(B > x)/\mathbf{E}[B]$.

Recall that $P_s = \mathbf{P}(D > zB)$. Consequently, if B is exponentially distributed, we have the insensitivity (w.r.t. the distribution of D) result $P_s = 1/\rho$.

Stochastic ordering

The inequality $P_s \leq 1/\rho$ holds if B^* is stochastically dominated by B, and $P_s \geq 1/\rho$ vice versa. Since B^* being stochastically dominated by B is related to a low variability of B, we see again that more variability (this time in the service times) leads to a better system performance (i.e. higher P_s).

Exponential reneging

If we assume that D has an exponential distribution (and B a general distribution), we see that z is the solution of

$$\rho\beta^*(z\nu) = 1, \text{ with } \beta^*(s) = \mathbf{E}\left[\mathrm{e}^{-sB^*}\right]. \qquad (4.3)$$

In addition, we have the following remarkable expression for the complete fluid limit $z(t), t \geq 0$, if $z_0 = 0$:

PROPOSITION 4.3. *Suppose $\mathbf{P}(D > t) = \mathrm{e}^{-\nu t}$, that B is independent of D and that $z_0 = 0$. Then the unique solution of (3.3) is given by $z(t) = z(1 - \mathrm{e}^{-\nu t})$, with z the solution of Equation (4.3).*

PROOF. Recall that Equation (3.3) has a unique solution. We show that $z(t)$ defined above is indeed the solution of (3.3) by verification. We thus compute the right hand side of (3.3) writing $z(u) = z(1 - \mathrm{e}^{-\nu u})$.

Observe that

$$z\int_s^t (1/z(u))du = \log(\mathrm{e}^{\nu t} - 1) - \log(\mathrm{e}^{\nu s} - 1).$$

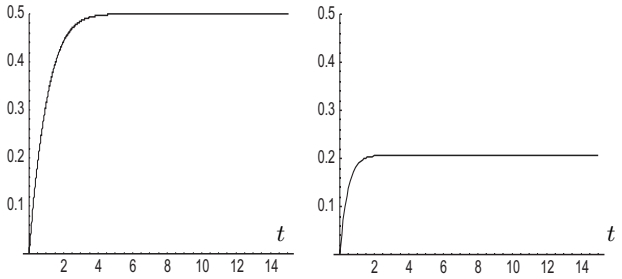

Figure 6: Fluid limit for B exponential and D exponential or Pareto

Consequently,

$$\begin{aligned}
&\lambda \int_0^t \mathbf{P}\left(D > t-s\right) \mathbf{P}\left(B > \int_s^t (1/z(u))du\right) ds \\
&= \frac{\lambda}{\nu} \mathrm{e}^{-\nu t} \int_0^t \mathbf{P}\left(zB > \log(\mathrm{e}^{\nu t}-1) - \log(\mathrm{e}^{\nu s}-1)\right) d\mathrm{e}^{\nu s} \\
&= \frac{\lambda}{\nu} \mathrm{e}^{-\nu t} \int_0^{\mathrm{e}^{\nu t}-1} \mathbf{P}\left(z\nu B > \log(\mathrm{e}^{\nu t}-1) - \log y\right) \mathrm{d}y \\
&= \frac{\lambda}{\nu} \mathrm{e}^{-\nu t} \int_{-\log(\mathrm{e}^{\nu t}-1)}^{\infty} \mathrm{e}^{-v} \mathbf{P}\left(z\nu B > \log(\mathrm{e}^{\nu t}-1) + z\right) \mathrm{d}v \\
&= \frac{\lambda}{\nu} \mathrm{e}^{-\nu t} (\mathrm{e}^{\nu t}-1) \int_0^{\infty} \mathbf{P}\left(z\nu B > v\right) \mathrm{e}^{-v} \mathrm{d}v \\
&= z(1-\mathrm{e}^{-\nu t})\rho \int_0^{\infty} \frac{\mathbf{P}\left(B > v/z\nu\right)}{\mathbf{E}\left[B\right]} \mathrm{e}^{-z\nu \frac{v}{z\nu}} \mathrm{d}(v/z\nu) \\
&= z(1-\mathrm{e}^{-\nu t})\rho\beta^*(z\nu) = z(1-\mathrm{e}^{-\nu t}).
\end{aligned}$$

Which shows that $z(1-\mathrm{e}^{-\nu t})$ satisfies (3.3). □

In general, it is impossible to compute $z(t)$ exactly, but again, Picard iteration turns out to be a quite tractable procedure. In Figures 6 and 7 below, we compute the fluid limit $z(t)$ for several different cases using Picard iteration. In all cases, $\lambda = 3/4$, B has either an exponential distribution or a Pareto distribution (as given before) with mean 2, and D is also exponential or Pareto with mean 2, independent of B. The cases where D are exponentially distributed are included for comparison purposes.

From the pictures, it appears that the speed of convergence towards steady state is exponentially fast, if either B or D has an exponentially bounded tail. We were not able to make this rigorous. Another striking fact is the difference of the limiting values which are respectively given by 0.5, 0.2067, 0.1174 and 0.0505. We see that more variability in the service and/or reneging time distribution has quite a positive impact on the performance of the system. Another feature we observe is that the limiting value of $z(t)$ is much lower than in the case where B and D are positively dependent with the same means. This is obvious since removing such dependence between B and D increases the fraction of early departures.

4.3 TCP-friendly traffic

To illustrate the versatility of our fluid model, assume that there exist independent random variables B_1 and D_1 with finite means such that

$$\begin{aligned}
(B, D) &= (B_1, \infty) && \text{with probability } p, \\
&= (\infty, D_1) && \text{with probability } 1-p.
\end{aligned}$$

This models the integration of elastic (TCP) traffic and TCP friendly UDP traffic. Key *et al.* [17] consider a related model in a network setting, but assume that all underlying random variables have exponential distributions. The fixed point equation (4.1) for z specializes to

$$z = \lambda p \mathbf{E}\left[zB_1\right] + \lambda(1-p)\mathbf{E}\left[D_1\right],$$

Consequently, if the stability condition $\lambda p \mathbf{E}\left[B_1\right] < 1$ is satisfied, we see that

$$z = \frac{\lambda(1-p)\mathbf{E}\left[D_1\right]}{1-\lambda p \mathbf{E}\left[B_1\right]}.$$

5. ADMISSION CONTROL

The numerous examples in the previous section showed that reneging has quite a negative impact in PS queues under overload. This raises the question of how to deal with this issue. Yang and De Veciana [26] propose to use a size based scheduling discipline like Shortest Remaining Processing time rather than Processor Sharing. Although these authors show by simulation that this leads to a better performance of the system in general, large jobs can significantly suffer from this change in policy. In addition, it has recently been shown by Verloop *et al.* [24] that the performance of size-based scheduling disciplines may be significantly worse (the stability region is even reduced) than fair sharing when one considers a network instead of a single link. In this section we do not aim to take a point of view about whether size based scheduling is better than fair sharing or not. Instead, we investigate another way of dealing with impatience and that is by introducing admission control, as suggested in [5].

Within the context of our fluid model, the simplest way to introduce admission control is to assume that the total mass in the system is bounded K, i.e. in the rth PSI queue, at most rK customers are allowed to be in the system simultaneously. To evaluate the steady state behavior of this model extension, we confine ourselves to a heuristic analysis as in Section 4. Assume that $\rho > 1$. Let q_K be the probability that a customer gets accepted upon arrival and let z_K be the number of customers in the system in steady state, both on fluid scale. As before, by Little's law, we see that z_K should satisfy the fixed-point equation

$$z_K = \lambda q_K \mathbf{E}\left[\min\{z_K B, I\}\right]$$

To solve this equation, one must know q_K. q_K can be seen as the limit of $\mathbf{P}\left(Z_K^r < rK\right)$ as $r \to \infty$. Assuming that Z_K^r/r converges a.s., two cases can occur: If $Z_K^r/r \to z_K < K$, then $\mathbf{P}\left(Z_K^r = K\right) \to 0$. In this case $q_K = 1$. Thus, z as defined before is a solution of z_K provided z is smaller than K, otherwise $z_K = K$. We conclude that $z_K = \min\{z, K\}$.

This results in an equation for q_K. If $z < K$ then $q_K = 1$ and by combining the above formulas we get

$$q_K = \frac{1}{\lambda \mathbf{E}\left[\min\{B, I/K\}\right]},$$

if $z \geq K$. The fraction of customers entering the system that leave successfully is given by $P_{s,K} = \mathbf{P}\left(z_K B < D\right)$. The total fraction of customers that get through successfully is then given by $V_K = P_{s,K} q_K$. To summarize, the fraction of successful customers V_K is given by

$$\begin{aligned}
V_K &= \frac{\mathbf{P}\left(KB < D\right)}{\lambda \mathbf{E}\left[\min\{B, D/K\}\right]}, && \text{if } K < z, \\
&= \mathbf{P}\left(zB < D\right) && \text{if } K \geq z.
\end{aligned}$$

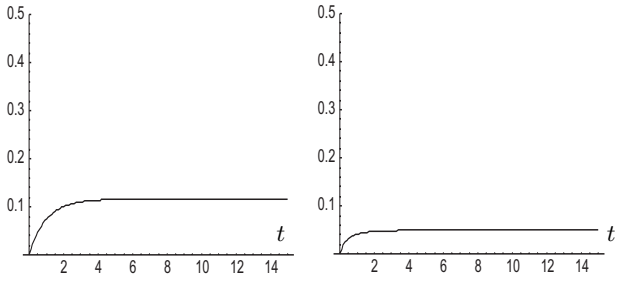

Figure 7: Fluid limit for B Pareto and D exponential or Pareto

It is easy to see that $V_K \to 1/\rho$ if $K \to 0$. The goodput, given by $\lambda q_K \mathbf{E}[B; B < D/z_K]$, converges to 1 as $K \to 0$, implying that the server will be completely utilized if the buffer size in the rth system is $o(r)$. From this point of view, it thus seems reasonable to perform admission control.

When we want to optimize the system by maximizing the fraction of customers that can be served successfully, things are not so simple. As in section 3, we both consider the example where B and D are linearly dependent and completely independent.

5.1 Linear dependence

We take the same setting as in Section 3.1: $D = \Theta B$, with Θ and B independent In this case V_K simplifies to

$$\begin{aligned} V_K &= \frac{\mathbf{P}(K < \Theta)}{\rho \mathbf{E}[\min\{1, \Theta/K\}]}, && \text{if } K < z, \\ &= \mathbf{P}(z < \Theta) && \text{if } K \geq z. \end{aligned}$$

For $K < z$, V_K can be written as

$$V_K = \frac{1}{\rho} \frac{\mathbf{P}(\Theta > K)}{\frac{1}{K}\int_0^K \mathbf{P}(\Theta > x)\, dx}.$$

From this representation, we see that V_K is decreasing in K on $[0, z]$ and we conclude that

PROPOSITION 5.1. *If $D = \Theta B$, then the number of successful transfers is asymptotically optimal as $r \to \infty$ if the buffer size K_r in the rth PSI queue satisfies $K_r \to \infty$, $K_r = o(r)$.*

Consequently, in the completely dependent case, optimizing the goodput also optimizes the fraction of successful customers.

5.2 Independent service times and lead times

In this case we get $V_K = \mathbf{P}(KB < D)/(\rho \mathbf{P}(KB^* < D))$ if $K < z$, and $\mathbf{P}(zB < D)$ otherwise. We immediately see that, if B has an exponential distribution, it does not matter for the users whether admission control is performed or not: $V_K = 1/\rho$ for all values of K (in particular for $K > z$ as shown in Section 4.2).

When we take the Pareto distribution $\mathbf{P}(B > x) = (a/(a+x))^b$ and also assume that D is a constant, then it can be shown that V_K is increasing in K for $K < z$, which implies that the best thing to do is to perform no admission control at all. If we take $\mathbf{P}(B > x) = x^{-b}$, it can be shown that V_K is maximized by $K = D$ if D is a constant.

These results seem somewhat surprising. The intuition is that our admission control policy does not discriminate between job sizes. Since the Pareto distribution generates relatively many small jobs, and some excessively large jobs, the impact of reneging is limited. We conclude by noting that the independence between B and D is probably not very realistic. In addition there exist more sophisticated admission control schemes which should lead to a better system performance in this case as well; see e.g. [2] for such an admission control scheme.

6. EXTENSIONS

This section briefly discusses how our model can be extended to deal with a number of other phenomena. We first discuss a more detailed description of user impatience in Section 6.1. After that, we discuss reattempts in Section 6.2.

6.1 More complex user behavior

Yang and De Veciana [26] consider a way to model user impatience involving both a global criterion (an upper bound on the total sojourn time) as a local criterion (a lower bound on the service rate). If it is clear that one of these criteria will not be met, the user reneges. The model we consider in the present paper only involves the global criterion, but can be extended to include local reneging behavior by defining U as the minimal instantaneous service rate a customer wishes to obtain. If we apply a similar scaling procedure, and observe the system in steady state, we see that the service rate remains the same, namely $1/z$, throughout the sojourn of a customer. Therefore, on fluid scale, a job that does not gets its instantaneous service rate leaves the system immediately. This effectively means that the service rate is reduced from λ to $\lambda \mathbf{P}(U < 1/z)$ and by applying Little's law, we see that the fixed point equation for z becomes

$$z = \lambda \mathbf{E}[\min\{zB, D\}(Uz < 1)].$$

It can be shown that this equation has a unique strictly positive solution if $\rho > 1$ and some weak regularity assumptions; we omit the details. The equation for $z(t)$ becomes (we assume $z(0) = 0$ for simplicity)

$$z(t) = \lambda \int_0^t \mathbf{P}(D > t - s; U < M(s,t); B > S(s,t))\, ds$$

with $M(s,t) = \min_{u \in [s,t]} 1/z(u)$. It would be interesting to investigate this functional equation in more detail.

6.2 Reattempts

Bonald & Roberts [5] propose to model user reattempts by assuming that a user leaving the system impatiently immediately reattempts with probability $p \in (0,1)$ with its initial value for B and D. If we follow their suggestion in our setting we obtain the following fixed-point equation for z. Let d be the stationary reneging rate. Then, applying Little's law, the equation for z becomes

$$\begin{aligned} z &= \lambda \mathbf{E}[\min\{zB, D\}] + dp \mathbf{E}[\min\{zB, D\} \mid zB > D] \\ &= \lambda \mathbf{E}[\min\{zB, D\}] + dp \mathbf{E}[D \mid zB > D], \end{aligned}$$

it is clear that due to our assumptions, a customer reneging once, will renege again later on. The stationary reneging rate d is therefore given by the equation $d = \lambda \mathbf{P}(zB > D) + pd$, i.e. $d = \lambda \mathbf{P}(zB > D)/(1-p)$. Combining the equations for z and d we see that

$$z = \lambda \mathbf{E}[\min\{zB, D\}] + \lambda p \mathbf{E}[D(zB > D)]/(1-p).$$

What we find interesting about this equation is that it may even have a strictly positive solution if $\rho < 1$. We conjecture therefore that reattempts may cause metastability: The system is not very congested for a long time, but due to some rare event, the queue length blows up, after which the reneging rate becomes so large that the effective load, due to reattempts, becomes structurally larger than 1. We intend to investigate this phenomenon in a future study.

7. CONCLUSION

We have considered a processor sharing queue with impatient customers. As this model is far too difficult to analyze exactly we have proposed a scaling procedure which leads to a tractable fluid approximation. We have used this fluid approximation to analyze the performance of the PSI queue in overload. As expected from earlier work, we have found that user impatience has quite a significant negative impact on system performance, also on finite time scales. By various stochastic ordering results we have shown that more variability has a positive impact on system performance. Finally, we have investigated the potential of admission control to control the negative effects of user impatience. We have shown that the suggested admission control reduces the impact of reneging on system performance in some cases. The fluid model considered here may be extended to include features as more complex reneging behavior and reattempts; we think that both extensions are an interesting topic for future research.

8. REFERENCES

[1] D. Barrer, *Queueing with impatient customers and ordered service*, Operations Research **5** (1957), 650–656.

[2] N. Benameur, S. Ben Fredj, S. Oueslati-Boulahia, and J.W. Roberts, *Quality of service and flow level admission control in the internet*, Computer Networks **40** (2002), 57–71.

[3] Thomas Bonald and Alexandre Proutière, *Insensitive bandwidth sharing in data networks*, Queueing Syst. **44** (2003), no. 1, 69–100.

[4] ———, *On stochastic bounds for monotonic processor sharing networks*, Queueing Syst. **47** (2004), no. 1-2, 81–106.

[5] Thomas Bonald and James Roberts, *Congestion at flow level and the impact of user behaviour*, Computer Networks **42** (2003), 521–536.

[6] N. Boots and Tijms H., *A multi-server queueuing system with impatient customers*, Management Science **45** (1999), 444–448.

[7] E. Coffman, A. Puhalskii, M. Reiman, and P. Wright, *Processor shared buffers with reneging*, Performance Evaluation **19** (1994), 25–46.

[8] Bogdan Doytchinov, John Lehoczky, and Steven Shreve, *Real-time queues in heavy traffic with earliest-deadline-first queue discipline*, Annals of Applied Probability **11** (2001), no. 2, 332–378.

[9] N. Gans, G. Koole, and A. Mandelbaum, *Telephone call centers: Tutorial, review, and research prospects*, Manufacturing & Service Operations Management **5** (2002), 79–141.

[10] Christian Gromoll, Philippe Robert, and Bert Zwart, *Fluid limits for processor sharing queues with impatience*, February 2006, http://www-rocq.inria.fr/˜robert/src/papers/2006-2.pdf.

[11] H. C. Gromoll, Amber Puha, and R. J. Williams, *The fluid limit of a heavily loaded processor sharing queue*, Annals of Applied Probability **12** (2002), no. 3, 797–859.

[12] Fabrice Guillemin, Philippe Robert, and Bert Zwart, *Tail asymptotics for processor-sharing queues*, Advances in Applied Probability **36** (2004), 525–543.

[13] Alain Jean-Marie and Philippe Robert, *On the transient behavior of some single server queues*, Queueing Systems, Theory and Applications **17** (1994), 129–136.

[14] W. Kang, F. Kelly, N. Lee, and R. Williams, *Fluid and brownian approximations for an internet congestion control model*, Proceedings of the 43rd IEEE Conference on Decision and Control, 2004, pp. 3938–3943.

[15] F. Kelly, *Charging and rate control for elastic traffic*, European Transactions on Telecommunications **8** (1997), 33–37.

[16] F. P. Kelly and R. J. Williams, *Fluid model for a network operating under a fair bandwidth-sharing policy*, Ann. Appl. Probab. **14** (2004), no. 3, 1055–1083.

[17] P. Key, L. Massoulié, A. Bain, and F. Kelly, *Fair internet traffic integration: Network flow models and analysis*, Annals of Telecommunications **59** (2004), 1338–1352.

[18] Lukasz Kruk, John Lehoczky, Steven Shreve, and Shu-Ngai Yeung, *Multiple-input heavy-traffic real-time queues*, Annals of Applied Probability **13** (2003), no. 1, 54–99.

[19] Laurent Massoulié and James Roberts, *Bandwidth sharing: Objectives and algorithms*, INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, 1999, pp. 1395–1403.

[20] J.. Mo and J. Walrand, *Fair end-to-end window-based congestion control*, IEEE/ACM Transactions on Networking **8** (2000), 556–567.

[21] A. L. Puha, A. L. Stolyar, and R. J. Williams, *The fluid limit of an overloaded processor sharing queue*, Preprint, 2004.

[22] Robert E. Stanford, *Reneging phenomena in single channel queues*, Mathematics of Operations Research **4** (1979), 162–178.

[23] ———, *On queues with impatience*, Advances in Applied Probability **22** (1990), no. 3, 768–769.

[24] M. Verloop, S. Borst, and R Nunez-Queija, *Stability of size-based scheduling disciplines in resource-sharing networks*, Performance Evaluation **62** (2005), 247–262.

[25] A. Ward and P. Glynn, *A diffusion approximation for a markovian queue with reneging*, Queueing Systems **43** (2003), 103–128.

[26] S.-C. Yang and G. de Veciana, *Bandwidth sharing: The role of user impatience*, GLOBECOM '01, 2001, pp. 2258–2262.