Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

Monotonicity in the limited processor sharing queue

M. Nuyens, W. van der Weij

REPORT PNA-E0802 JUNE 2008

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

Monotonicity in the limited processor sharing queue

ABSTRACT
We study a processor sharing queue with a limited number of service positions and an infinite buffer. The occupied service positions share an underlying resource. We prove that for service times with a decreasing failure rate, the queue length is stochastically decreasing in the number of service positions, and that for service times with an increasing failure rate, the queue length is stochastically increasing. The result is illustrated with simulations, and the queue length is compared to that in other queueing models with and without restrictions on the number of service positions.

2000 Mathematics Subject Classification: 68M20;60K25;90B36
Keywords and Phrases: Monotonicity; LPS queue; Limited Processor Sharing;

Monotonicity in the limited processor sharing queue

Misja Nuyens* and Wemke van der Weij†

26th May 2008

Abstract

We study a processor sharing queue with a limited number of service positions and an infinite buffer. The occupied service positions share an underlying resource. We prove that for service times with a decreasing failure rate, the queue length is stochastically decreasing in the number of service positions, and that for service times with an increasing failure rate, the queue length is stochastically increasing. The result is illustrated with simulations, and the queue length is compared to that in other queueing models with and without restrictions on the number of service positions.

1 Introduction

When the First Come First Served (FCFS) discipline was superseded in popularity for a number of applications, the Processor Sharing discipline (PS) was one of the main disciplines to take over. Not only does PS perform much better under heavy-tailed service-time distributions, but it is also relatively easy to implement, as no information on the job size

*Statkraft Markets, Amsterdam, The Netherlands, misjanuyens@gmail.com

†CWI, Amsterdam, The Netherlands, w.van.der.weij@cwi.nl

is needed. Furthermore, PS is in many ways a fair discipline, since it does not discriminate among jobs based on their arrival time, original size or remaining size. See, for example, Wierman and Harchol-Balter [13].

However, the PS scheduling mechanism is not always feasible in practice: although the number of jobs in the system may be unbounded, the number of jobs being served simultaneously may not. A typical example is found in modeling web servers: these are equipped with a (finite) number of so-called threads than can handle incoming web transactions requests. A specific feature of these threads is that they effectively share an underlying hardware resource [4, 7].

To overcome the infeasibility of the ordinary PS model, we discuss the so-called *limited processor sharing discipline with c service positions* (LPS-c, or shortly, LPS) and an infinite buffer. In the LPS-c queue, at most c jobs can receive service simultaneously. If there are more jobs in the queue, they have to wait until one of the c jobs in service leaves. Jobs are always accepted in the queue, since the queue (or buffer) has infinite capacity. So, when there are k jobs in the queue, the service rate for the jobs in service is $1/\min\{k, c\}$. Clearly, LPS-1 is the same as FCFS, and in the limit $c \to \infty$, LPS-c is equal to ordinary PS.

Apart from these limiting scenarios, and the case of exponential service times (where the queue-length distribution does not depend on the service discipline), little is known about the LPS queue. Avi-Itzhak and Halfin [1] provide some preliminary insights for the general LPS system, and give an approximation for the expected sojourn time. Unfortunately, this approximation is only accurate when the coefficient of variation is small. By simulations, van der Weij [12] obtained some insights in the behaviour of the LPS model for small values of c, pointing in the direction that the expected sojourn time could be monotone in c. Very recently, Zhang et al. [14, 15] and Zhang and Zwart [16] have investigated the LPS

queue, and described the behaviour of the queue in light and heavy traffic. They derived an approximation for the waiting probability in the limit $c \to \infty$ [14, 15], and for the steady-state queue length and response time in heavy traffic [16].

The main result of this paper is that for a class of service-time distributions, namely distributions with a decreasing failure rate, the queue length in the LPS queue is monotonically decreasing in c, in the stochastic order sense. For distributions with an increasing failure rate, the reverse statement holds. Examples of distributions with a decreasing failure rate are Pareto distributions, and (certain) Gamma and Weibull distributions. The normal distribution (at the non-negative domain) and the uniform distribution have an increasing failure rate.

In certain applications where the service can be preemptive, it is the number of preempted jobs and jobs in service that may be limited, rather than the number of service positions. For this model, we introduce the limited Foreground-Background (LFB) queue. The FB discipline is a well-known discipline and minimises the expected queue length in an $G/GI/1$ queue for a service-time distribution with decreasing failure rate, see Righter and Shanthikumar [11]. Restricting the number of preempted jobs and jobs in service c then yields the LFB-c model. The proof used for showing monotonicity in the number of service positions for the LPS queue can be used to prove a similar result for the LFB-c queue.

The paper is organised as follows. In Section 2, we introduce the model and notation. The main result is proved in Section 3. It is followed by a number of corollaries, and illustrated by simulations. The results are extended to the LFB queue in Section 4. In Section 5, we briefly discuss the performance of the LPS queue under another performance metric: the tail of the sojourn time distribution. We conclude in Section 6 by providing directions for further research.

2 Model and preliminaries

We consider a queueing model with one queue. Jobs arrive according to an unspecified process. We model the service times by a non-negative continuous random variable with distribution function $F(x)$ and density function $f(x)$. The queue has an infinite buffer but a finite number of service positions, denoted by c. As long as there are c or less jobs in the system, the queue operates as under the ordinary PS queue, but as soon as the number of jobs in the queue exceeds c, the behaviour of the queue becomes different: the $(c+1)$st job has to wait until one of the c jobs that are in service leaves the system. The service is non-preemptive. This model is denoted by $G/GI/1$ LPS-c (or LPS).

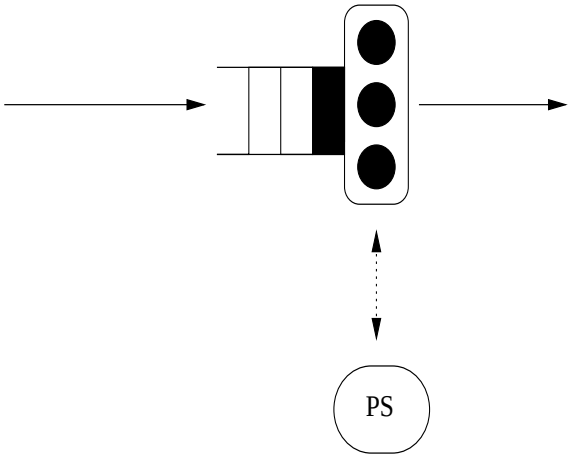

Figure 1: Illustration of the LPS-c model for $c = 3$.

For the service-time distributions, we focus on two cases, namely service-time distributions with an increasing failure rate (IFR) and with a decreasing failure rate (DFR), defined as follows. The failure (or hazard) rate, denoted by $\mu(x)$, is defined as:

$$\mu(x) = \frac{f(x)}{1 - F(x)}, \qquad x \geq 0. \tag{1}$$

A service-time distribution belongs to the class DFR if $\mu(x)$ is decreasing for all x, i.e., $\mu(x) \geq \mu(y)$ whenever $x \leq y$. A service-time distribution belongs to the class IFR if $\mu(x)$ is increasing for all x.

Finally, a random variable X is said to be *stochastically smaller* than a random variable Y, notation $X \leq_{st} Y$, if $P(X > x) \leq P(Y > x)$ for all x.

3 Monotonicity results

We are now ready to prove the main result of the paper: a characterisation of the behaviour of the queue length in the $G/GI/1$ LPS-c queue with respect to c, the number of service positions.

Theorem 1 *Let $X^c(t)$ denote the queue length in the $G/GI/1$ LPS-c queue at time t. If the service times have a DFR distribution, then for all $t \geq 0$ and $c \in \{1, 2, \ldots\}$, $X^{c+1}(t) \leq_{st} X^c(t)$. For IFR distributions, the stochastic inequality is reversed.*

Proof As in Righter and Shanthikumar [11], we construct the proof for a discrete-time process. In discrete time, the LPS-c discipline serves the jobs (at most c) in a round-robin fashion. To prove the result, we use artificial disciplines denoted by LPS-$c\Box n$, for $n = 0, 1, \ldots$, defined as follows. The first n time steps, LPS-$c\Box n$ behaves exactly like the LPS-$(c+1)$ discipline. From time step $n+1$ onwards, it behaves like the LPS-c discipline. So, intuitively one can think of LPS-$c\Box n$ as a mixture of LPS-c and LPS-$(c+1)$. Furthermore, if there are not more than c jobs in the system up to time n, then LPS-$c\Box n$ and LPS-c are identical. The core of the proof is showing that for all $c \in \mathbb{N}$, and $n \in \{0, 1, 2, \ldots\}$,

$$X^{c\Box n}(t) \geq_{st} X^{c\Box n+1}(t), \qquad t \geq 0. \tag{2}$$

The theorem then follows from noting that $X^c(t) = X^{c\square 0}(t)$ and $X^{c+1}(t) = \lim_{n\to\infty} X^{c\square n}(t)$ for all $t \geq 0$.

To prove (2), fix a $t > n$ (for $t \leq n$ there is nothing to prove) and note that LPS-$c\square n$ and LPS-$c\square(n+1)$ are the same up to time n. If LPS-$c\square n$ and LPS-$c\square(n+1)$ also serve the same job at time $n+1$, then they are equal forever, and there is nothing left to prove. So, assume that LPS-$c\square n$ and LPS-$c\square(n+1)$ serve a different job at time $n+1$. Denote the job served by LPS-$c\square(n+1)$ at time $n+1$ by x, and let $a(x)$ denote the amount of service it has received by time $n+1$. After that, until job x leaves the queue, LPS-$c\square(n+1)$ serves at each time step the job that was served by LPS-$c\square n$ at the previous time step.

Now there are two cases. First, if LPS-$c\square n$ serves job x at a time $u \leq t$ (note that this is only possible if one of the jobs leaves the LPS-$c\square n$ queue before time t), then at time step u, job x has received the same amount of service under both policies, as have all other jobs. Hence, the queues are the same at time u, and from that moment on, the two queue lengths will be identical.

To conclude the proof, we consider the case that job x is not served by the LPS-$c\square n$ queue before or at time t. Hence, at time t there are exactly two jobs that have received different amounts of service under LPS-$c\square n$ and LPS-$c\square(n+1)$: job x and the job served by LPS-$c\square n$ at time t, denoted by y, with received service $a(y)$, see also the table below.

time	1	$\cdots$	n	$n+1$	$n+2$	$\cdots$	$t-1$	t
job served by PS-$c\square n$	same	$\cdots$	same	d	e	$\cdots$	g	y
job served by PS-$c\square n+1$	job	$\cdots$	job	x	d	$\cdots$	f	g

As a consequence, the queue lengths in the two queues can differ by at most 1. The crucial

observation is now that $a(y) \geq a(x)$. Since μ is decreasing by assumption, this implies that

$$\begin{aligned}
\mathbb{P}(X^{c\Box n}(t) = X^{c\Box n+1}(t) - 1) &= \mathbb{P}(\text{job } y \text{ leaves at time } t, \text{ job } x \text{ has not left at time } n+1) \\
&= \mu(a(y))[1 - \mu(a(x))] \\
&\leq \mu(a(x))[1 - \mu(a(y))] \\
&= \mathbb{P}(\text{job } x \text{ has left at time } n+1, \text{ job } y \text{ does not leave at time } t) \\
&= \mathbb{P}(X^{c\Box n}(t) = X^{c\Box n+1}(t) + 1). \qquad (3)
\end{aligned}$$

Since $X^{c\Box n}(t)$ and $X^{c\Box n+1}(t)$ can differ at most 1, we conclude from (3) that $X^{c\Box n}(t) \geq_{st} X^{c\Box n+1}(t)$. For IFR service-time distributions, the inequality in (3) is reversed. This completes the proof. □

Theorem 1 matches with the following heuristic. For DFR service times, among all work-conserving disciplines P, the queue length $X(t)^P$ is (stochastically) maximal under P=FCFS, and minimal under FB, for all t, see Righter and Shanthikumar [11]. For $c = 1$, LPS is the same as FCFS. Furthermore, the larger c, the more LPS differs from FCFS, and the more it behaves like FB. Hence, intuitively, the larger c, the smaller the queue length.

Assuming that the load ρ satisfies $\rho < 1$, the workload process converges to a stationary state as the time goes to ∞, see Kelly [10]. Let X^c be the stationary queue length in the LPS-c queue. By letting the time t go to ∞, we obtain the following corollary.

Corollary 2 *If the service times have a DFR distribution, and $\rho < 1$, then $X^{c+1} \leq_{st} X^c$ for all $c \in \mathbb{N}$. For IFR service times, $X^{c+1} \geq_{st} X^c$ for all $c \in \mathbb{N}$.*

It is interesting to compare Corollary 2 with the following heavy-traffic approximation of the expected queue length found in Zhang and Zwart [16]:

$$\mathbb{E}X^c \approx \frac{\rho}{1-\rho}\frac{\nu_a^2+\nu_s^2}{2(1+\nu_s^2)}\Big(2+(\nu_s^2-1)\rho^{c\frac{1+\nu_a^2}{\nu_a^2+\nu_s^2}}\Big),$$

where ν_a and ν_s are the coefficients of variation of the interarrival and service-time distributions. This expression is decreasing in c if and only if $\nu_s \geq 1$. Since all DFR distributions satisfy this condition, the approximation of $\mathbb{E}X$ is decreasing in c for a class of distributions that includes DFR distributions. On the other hand, the ordering in Corollary 2 is a stronger, namely stochastic in stead of 'in expectation'. Obviously, for IFR distributions, comparable statements hold.

Little's law implies the following result for the stationary sojourn time, S^c.

Corollary 3 *If the service times have a DFR distribution, and $\rho < 1$, then $\mathbb{E}S^{c+1} \leq \mathbb{E}S^c$. For IFR and $\rho < 1$, we have $\mathbb{E}S^{c+1} \geq \mathbb{E}S^c$.*

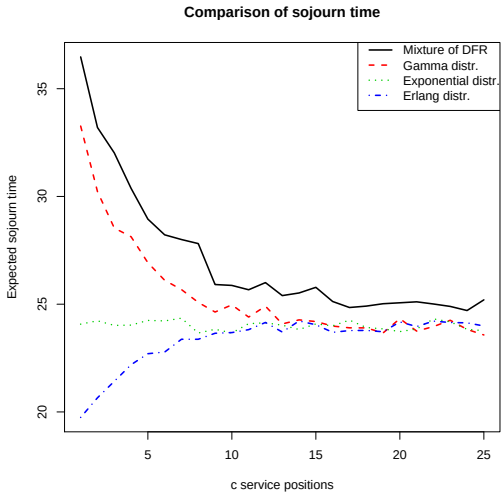

Figure 2: The expected sojourn time in the $M/GI/1$ LPSqueue for $\lambda = 4/15$, $\rho = 0.8$ and Erlang(2), exponential, Gamma(0.5, 12) and a mixture of DFR service-time distributions.

To illustrate Corollary 3, in Figure 2 we have simulated the expected sojourn time for a number of $M/GI/1$ LPS queues where the service times have strictly decreasing (Gamma and multi-class exponential), strictly increasing (Erlang) and constant failure rates (exponential).

Consider two density functions, f_1, and f_2. Let f be the mixture of these density functions, defined by $f = \alpha f_1 + (1-\alpha)f_2$, where $\alpha \in (0,1)$. Since a mixture of DFR distributions is again DFR, see Barlow and Proschan [2], Theorem 1 automatically also holds for multi-class queues where the distribution in each class is DFR, as stated in the following corollary.

Corollary 4 *Let $X^c(t)$ denote the queue length at time t in the $G/GI/1$ LPS queue, where jobs are allowed to come from different classes. If the service-time distribution in each class is DFR, then $X^{c+1}(t) \leq_{st} X^c(t)$ for all $t \geq 0$.*

The reverse does generally not hold for IFR distributions, since a mixture of IFR distributions is not necessarily IFR. For example, a mixture of exponential distributions is hypergeometric, which is not IFR. But a mixture of normal distributions is IFR, if they have the same variance and $|\mu_1 - \mu_2| \leq 2\sigma$ (where μ_i is the mean and σ^2 is the variance), see Block et al. [3].

The DFR condition in Theorem 1 is quite important. At first glance, one might think that it would be possible to weaken the condition on the service times from DFR to 'having a coefficient of variation larger than 1'. However, the simulations in Figure 3 contradict this. In Figure 3, we have simulated $\mathbb{E}(X^c)^2$ for a number of $M/GI/1$ LPS queues with a lognormal service-time distribution with coefficient of variation 1.72, mean job size 1 and $\rho = 0.8$, for different values of c. The failure rate of this distribution is first increasing,

and then decreasing. From the figure it is clear that $\mathbb{E}(X^c)^2$ is not monotone in c. Since $X \leq_{st} Y$ if and only if $\mathbb{E}h(X) \leq \mathbb{E}h(Y)$ for all increasing functions h, we conclude that X^c is not stochastically monotone in c. Hence, for this distribution with coefficient of variation larger than 1, Theorem 1 does not hold.

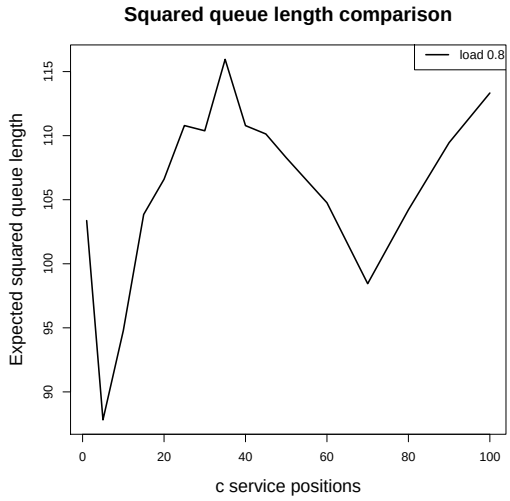

Figure 3: Simulation results for $\mathbb{E}(X^c)^2$ in the $M/GI/1$ LPS queue with lognormally distributed service times with mean 1 and squared coefficient of variation 1.72 for $\rho = 0.8$.

4 Other limited queues

In certain applications where the service may be preemptive, the limit is not so much on the number of service positions, but it is the number of preempted jobs that is limited, see for example van der Mei et al. [7]. Assume that at most c jobs are allowed to have received service. These jobs constitute the *inner queue*. All other jobs are waiting in the *outer queue*, and are only allowed to enter the inner queue when the number of jobs in the inner is smaller than c, see Figure 4. We call this queue the *inner-outer queue with c positions,* and denote it by $G/GI/1/c/\infty$. The LPS queue with c service positions discussed in the

previous section is an example of an inner-outer queue.

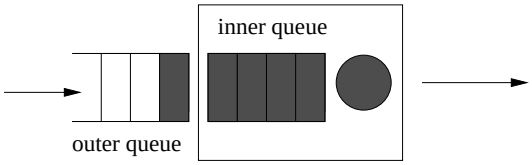

Figure 4: The $G/GI/1/c/\infty$ model.

In this section, we discuss the discipline that stochastically minimises the queue length in the $G/GI/1/c/\infty$ queue, where the queue length is the sum of the jobs in the inner queue plus the outer queue. We introduce the limited FB discipline, denoted by LFB, which minimises the queue length under DFR service-time distributions. Then, using the proof technique used in the previous section, we show that for DFR distributions, under the LFB discipline, the queue length is stochastically decreasing in c.

First, let us describe how the normal FB discipline works. Denoting by the *age* of a job the amount of service it has received, FB serves the youngest job. If there are n such jobs, they are served simultaneously, in a processor sharing manner, at rate $1/n$. In particular, when a new job arrives, the job in service is preempted, and the new job is served immediately. The preempted job is assumed to occupy a service position. See Nuyens and Wierman [8] for a recent survey on the FB discipline.

We now modify the FB discipline to fit to the $G/GI/1/c/\infty$ framework. The *limited* FB *queue* with c positions, denoted by LFB-c, is defined as follows: under LFB-c, the server preemptively serves the youngest job, but only if there are c jobs or less in the inner queue. If there are c jobs in the inner queue, the server can only switch to an unserved job when one of the c jobs with positive age leaves the queue. Like in the PS queue, LFB-1 is equal to FCFS, while for $c \to \infty$, LFB-c converges to the ordinary FB discipline. The FB discipline

can be considered to be the opposite of FCFS: FB always serves the youngest job(s), while FCFS serves the oldest. Hence, LFB-c has the interesting feature that when c runs from 1 to ∞, LFB-c moves from FCFS to its opposite, FB.

An important observation is that among all service disciplines in the $G/GI/1/c/\infty$ queue, LFB-c always serves as much new jobs as possible. Since it is exactly this property that makes the proof of the optimality of FB (Theorem 2.1 of Righter and Shantikumar [11]) work, we have the following result:

Theorem 5 *Consider the $G/GI/1/c/\infty$ queue with DFR service times. Let $X^{\mathsf{P}}(t)$ denote the queue length at time t under discipline P. Then for all policies P that do not use information on the exact job sizes, we have $X^{\mathsf{LFB}-c}(t) \leq_{st} X^{\mathsf{P}}(t) \leq_{st} X^{\mathsf{FCFS}}(t)$. For IFR job sizes, the inequalities are reversed.*

Copying the proof of Theorem 1 implies the following stochastic ordering of the different queue lengths discussed in this paper:

Theorem 6 *For the $G/GI/1/c/\infty$ LFB-c queue with a DFR service-time distribution, we have that $X^{c}(t) \geq X^{c+1}(t)$ for all t and all c. For IFR service times, the inequalities are reversed.*

Combining Theorems 1, 5 and 6 with Theorem 2.1 of Righter and Shantikumar [11], we have the following result.

Corollary 7 *In the $G/GI/1$ queue with DFR service times, we have for $\P \in \{LFB-c, \mathsf{PS}\}$:*

$$X^{\mathsf{FB}} \leq_{st} X^{\P} \leq_{st} X^{\mathsf{LPS}-c} \leq_{st} X^{\mathsf{FCFS}}.$$

For IFR service times, the inequalities are reversed.

5 The tail of the sojourn-time distribution

In the previous section, we have seen that for IFR service times, the performance of the LPS queue is decreasing in c if the queue length is used as a performance measure. In this section, we show that for a class of light-tailed service distributions, which contains most IFR distributions, the behaviour of LPS in heavy-traffic under another performance measure is optimal.

This other performance measure indicates the likelihood of a very long sojourn time, i.e., the total time spent in the system. For light-tailed distributions, i.e., distributions for which there exists an $s > 0$ such that $\mathbb{E}[\exp(sB)] < \infty$, the tail of the distribution can be characterised by its *decay rate*. The decay rate γ of a random variable U is defined as

$$\gamma(U) = -\lim_{z\to\infty} \frac{1}{u} \log P(U > u).$$

Hence, the smaller the decay rate, the larger the tail of the distribution. Denoting the sojourn time of a discipline P by S_{P}, we prove the following theorem.

Theorem 8 *In the $G/GI/1$ queue, if $\mathbb{E}[\exp(sB)] < \infty$, then for all c, there exists a $\rho(c) < 1$ such that $\gamma(S_{\mathsf{LPS}}) = \gamma(S_{\mathsf{FCFS}})$ for all $\rho \geq \rho(c)$.*

Before giving the proof, we discuss the result. In case of light-tailed service times, the decay rate of the sojourn time under any work-conserving discipline lies in the (non-empty) interval $[\gamma(L), \gamma(W)]$, where W is the stationary workload, and L the length of a generic busy period. Most well-known policies have a decay rate that equals one of these extremes. The lower bound is matched for LCFS, FB, and, under mild additional conditions, SRPT, and PS. Under FCFS, the upper bound is achieved. For an overview of these results on the decay rates of different disciplines, see Nuyens and Zwart [9] and the references therein.

The remarkable conclusion we can draw from Theorem 8 is that no matter how large c is, the decay rate of the sojourn time under LPS in heavy traffic is equal to that of FCFS, which is optimal, and not to that of PS, which in most cases is the worst possible. This means that in heavy traffic, the "FCFS-like" property of LPS that once a job is served, it is guaranteed a minimum rate, becomes much more important than the "PS-like" property that the service rate is shared with other jobs in the system.

Proof of Theorem 8 The sojourn time of a job of size B in the LPS queue can be upper bounded by $W + cB$, where W and B are independent: the job has to wait at most W before its service starts, and then it is served with rate at least $1/c$. Since $\gamma(X+Y) = \min\{\gamma(X), \gamma(Y)\}$ if X and Y are independent, see for instance [5], we have

$$\gamma(S_{\mathsf{PS}-c}) \geq \min\{\gamma(W), \gamma(B)/c\}.$$

Since $\gamma(W) \to 0$ as $\rho \to 1$, see for example Mandjes and Zwart [6], there exists a $\rho(c)$ such that $\gamma(W) \leq \gamma(B)/c$, and hence $\gamma(S_{\mathsf{LPS}}) \geq \gamma(W)$. The proof is finished by noting that $\gamma(S_{\mathsf{LPS}}) \leq \gamma(W)$ holds for all work-conserving disciplines, see also Nuyens and Zwart [9]. □

In some cases, $\gamma(W)$ and $\gamma(B)$ can be calculated explicitly. For example, in the $M/M/1$ queue where the interarrival and service times have parameters λ and μ, we have $\gamma(W) = \mu - \lambda$ and $\gamma(B) = \mu$. Writing $\rho = \lambda/\mu$, we have

$$\gamma(W) \leq \frac{\gamma(B)}{c} \;\leftrightarrow\; \mu - \lambda \leq \frac{\mu}{c} \;\leftrightarrow\; 1 - \rho \leq \frac{1}{c} \;\leftrightarrow\; \rho \geq 1 - \frac{1}{c}.$$

Hence, for all ρ such that $1 - c^{-1} \leq \rho < 1$, we have $\gamma(S_{\mathsf{LPS}}) = \gamma(W) = \mu - \lambda$.

6 Extensions and further research

In this section we discuss two possible extensions of the LPS model that are interesting for further research.

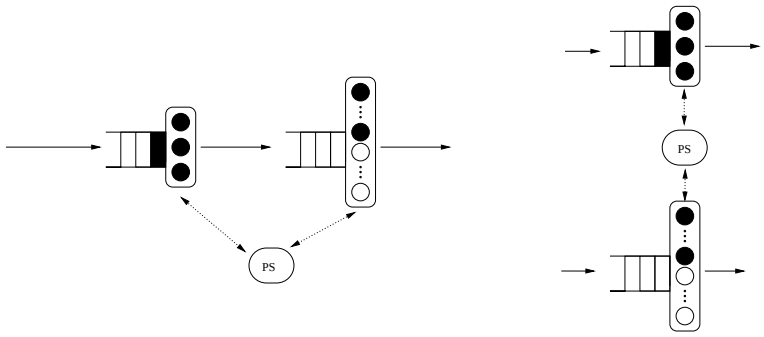

Figure 5: Possible extensions of the LPS model.

One possible extension of the LPS model is to consider two $G/GI/1$ LPS queues in tandem, see the left part of Figure 5. In this tandem model, jobs arrive according to an unspecified process at the first queue. After completion of service, they are routed to the second queue, and after the completion of the service at that queue the jobs leave the network. The queues have infinite buffers but finite numbers of service positions, denoted by c_1 and c_2. As long as there are c_i or less jobs in each queue (for $i = 1, 2$), the queue mimics the PS queue where the total capacity is equally shared among all occupied service positions at the two queues. But as soon as the number of jobs in queue i, say, exceeds c_i, the behaviour of system differs: the $(c_i + 1)$st job has to wait until one of the c_i jobs that are in service leaves the system.

For this queue, a research question is under which conditions there is monotonicity in c_1 and c_2. For exponentially distributed service times, the expected overall queue length is minimised if $c_1 = 1$ and c_2 as large as possible, while for each queue in isolation the

expected queue length is increasing in c_i, see also [12].

A second extension of the LPS model is the following. Consider again two $G/GI/1$ LPS queues, and assume that these queues share a common resource in the same manner as the model above. The two arrival streams are now independent, and each queue is fed by one of these arrival streams. After service at that queue, the job immediately leaves the system (instead of being routed to an other queue), see the right part of Figure 5. As far as we know, no monotonicity results on the queue length with respect to c_1 and c_2 are known. Since the number of service positions assigned to queue 1 influences the capacity assigned to queue 2, this model cannot be solved by the same techniques as used in this paper.

Acknowledgements The authors would like to thank Rhonda Righter, Michel Mandjes, and Bert Zwart for excellent comments on earlier version of this paper. Thanks!

References

[1] B. Avi-Itzhak and S. Halfin. Server sharing with a limited number of service positions and symmetric queues. *Journal of Applied Probability* **24**: 990-1000 (1987).

[2] R.E. Barlow and F. Proschan. Statistical theory of reliability and life testing. Probability models (1975).

[3] H.W. Block, Y. Li, T. H. Savits. Mixtures of normal distributions: Modality and failure rate. *Statistics and Probability Letters* **74**: 253-264 (2005).

[4] M. Harkema, B.M.M. Gijsen, R.D. van der Mei, and Y. Hoekstra. Middleware performance modelling. *Proceedings international Symposium on Performance Evaluation*

of Computer and Telecommunication Systems SPECTS San Jose, CA, 733-742, July (2004).

[5] M. Mandjes and M. Nuyens. Sojourn times in the M/G/1 FB queue with light-tailed service times. *Probability in the Engineering and Informational sciences* **19**(3): 351-361 (2005).

[6] M. Mandjes and B. Zwart. Large deviations for sojourn times in processor sharing queues. *Queueing Systems* **52**: 237-250 (2006).

[7] R.D. van der Mei, R. Hariharan and P.K. Reeser. Web server performance modeling. *Telecommunication Systems* **16**: 361-378 (2001).

[8] M. Nuyens and A. Wierman. The Foreground-Background queue: a survey. *Performance Evaluation* **65**: 286-307 (2008).

[9] M. Nuyens and B. Zwart. A large-deviations analysis of the GI/GI/1 SRPT queue. *Queueing Systems* **54**(2): 85-97 (2006).

[10] F.P. Kelly. Reversibility and Stochastic Networks. John Wiley & Sons, 1979.

[11] R. Righter and J.G. Shanthikumar. Scheduling multiclass single server queueing systems to stochastically maximize the number of successful departures. *Probability in the Engineering and Informational Sciences* **3**: 323-333 (1989).

[12] W. van der Weij. Sojourn times in a two-layered tandem queue with limited service positions and a shared processor, Master Thesis, University of Amsterdam (2004).

[13] A. Wierman and M. Harchol-Balter. Classifying scheduling policies with respect to unfairness in an M/GI/1. *Proceedings ACM Sigmetrics 2003 Conference on Measurement and Modeling of Computer Systems* San Diego, CA, June (2003).

[14] J. Zhang, J.G. Dai and B. Zwart. Diffusion limits of Limited Processor Sharing queues. *Submitted* (2008).

[15] J. Zhang, J.G. Dai and B. Zwart. Low of large number limits of Limited Processor Sharing queues. *Submitted* (2008).

[16] J. Zhang and B. Zwart. Steady state approximations of Limited Processor Sharing queues in heavy traffic. *Submitted* (2008)