The performance of a single server queue with preemptive random priorities

Moshe Haviv *

June 4, 2015

Abstract

Queues in which customers who belong to different classes, have different priority levels are an old subject. Usually one looks for the performance of each class given its priority level. We suggest here a new model. Specifically, we consider the M/G/1 queue model in which all customers are identical ex-ante but prior to joining the queue, they draw a random (preemptive) priority level. We derive the Laplace-Stieltjes transform (LST) of a customer given his drawn priority parameter. From that the LST of an arbitrary customer can be integrated out. We present a a number of proofs which give some insight to the model. Special attention in given to the case of exponential service (the M/M/1 queue) and to finding the first moment of waiting. In particular, we show that the model is 'a middle of the road one' in the sense that the mean sojourn time lies between the corresponding means under the FCFS and the Last come first served (LCFS-PR) (or equivalently, the Egalitarian Processor Sharing (EPS)) schemes. Finally, we show how the new scheme may lead to an improvement in the utilization of the server when customer decide whether or not to join. We conclude with a few words on the corresponding model but without preemption.

*Department of Statistics and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel. email:moshe.haviv@gmail.com.

1 Introduction

The purpose of this paper is to analyse a new queueing regime, that of preemptive random priority and derives its performance measures.[1] Specifically, under this regime any arrival draws his (absolute) priority parameter from a continuous distribution (as we will see, what is the actual distribution is immaterial). The one who receives service possesses the highest priority among those in the system. In particular, this is done with preemption. We concentrate on deriving the performance measures of this queueing regime under the standard model of an M/G/1 queue. We next state the needed model assumptions and the main results will appear in the sections which will follow.

The M/G/1 queueing model is one of the most researched model in operations research and in the performance evaluation area of computer sciences. In this model customers arrive to a single server queue with accordance to a Poisson process with rate denoted by λ per unit of time. Service times are independent and identically distributed with cumulative distribution function G. Denote by $G^*(s)$ the Laplace-Stieltjes transform (LST) of the service time, namely $G^*(s) = \mathrm{E}(e^{-sX})$ where $X \geq 0$ is a random variable representing a single service time. This distribution is not limited to belong to any specific family of distributions (such as the exponential family). Denote by $\overline{x}$ and $\overline{x^2}$ the first and second moments of service times, respectively. We denote $\lambda\overline{x}$ by ρ and assume for stability that $\rho < 1$. It is well-known that ρ is the server utilization level, namely the proportion of time where the server is busy.

The question of interest here is what is the distribution of the time in the system of an arbitrary customer (in steady-state conditions). This distribution is a function of the service regime. For example, this distribution is different if customers enter service in a first-come first-served (FCFS) order or in random order. Putting the pyramid on its head, one can select the service regime so as to achieve some goals in the resulting waiting time distribution, in particular, in its mean. As for the latter parameter, it stays the same under all non-preemptive, work-conserving regime.[2] and non-anticipating[3]

[1] This regime was introduced in [7] and this paper contains what appeared there coupled with the derivation of some additional performance measures.

[2] A queueing regime is said to be work-conserving if the total amount of work in the system at any give time is the same as in the corresponding FCFS case.

[3] A queueing regime is said to be non-anticipating if the decision who gets service at

Thus, in order to deviate from this common value, one needs to introduce preemption.

As said, we consider a new regime, that of preemptive random priority (PRP). Under this regime customers draw a random priority level from some (in fact, any) continuous distribution. Those who receive higher priority level enter first to service and, if needed, even preempt those with a lower level. The paper is devoted to finding the performance measures under this regime. We derive in Section 3.2 the Laplace-Stieltjes transform (LST) of a customer given his drawn priority parameter. From that the LST of an arbitrary customer can be integrated out. We present a number of proofs which give some insight into the model. Special attention in given in Section 3.1 to the case of exponential service time (the M/M/1 queue) and to finding the first moment of waiting in Section 2. In particular, we show that this mean time lies between the corresponding means under the FCFS and the Last Come First Served with Preemption Resume (LCFS-PR) (or equivalently, the Egalitarian Processor Sharing (EPS)) schemes. In Section 4 we show how this new regime can be incorporated into a to-queue-or-not-to-queue decision model and how it can lead to different resulting joining rate in comparison with the standard FCFS case. We conclude in Section 5 with some discussion concerning a similar regime but without preemption.

We are aware of two references in which the PRP regime is used. These are [2] and [4] (see also [5], pp.102-104). There, PRP is not assumed but rather turned out to be the resulting regime when customers have the option to pay in order to get a preemptive priority parameter. The more they pay, the higher is their preemptive priority level, reducing their mean waiting costs. In the latter reference, customers have also the option to opt out under a standard cost/reward assumption. Minding the trade-off between length of wait and the level of payment, and noticing that other customers face a similar dilemma, their equilibrium behavior is to pay a random amount (having some specific distribution) resulting overall in the PRP regime. In the latter model some will opt out and, interestingly, in the case of an exponential service time (i.e., an $M/M/1$ model), the fraction of those who join coincide with the socially optimal joining rate. The $M/M/1$ case is also dealt with in [8]. It is shown there that if the regime imposed is that of PRP and customers decide whether or not to join after drawing and inspecting their

any given time instant is not base on the actual (past or future) service times of those present in the system.

priority parameters, then the resulting equilibrium joining rate coincides with the socially optimal rate. Finally, the reader is referred to Part VII of [3] for a comprehensive summary on various queueing regimes.

2 Mean value analysis and comparison with well-known regimes

There are various possible service regimes. The most known one is First Come First Served (FCFS). Under this regime the mean waiting time in the system (queueing plus service), denoted by $\mathrm{E}(W_{FCFS})$, equals

$$\mathrm{E}(W_{FCFS}) = \frac{\lambda\overline{x^2}}{2(1-\rho)} + \overline{x}. \tag{1}$$

This is the well-known Khintchine-Pollaczek formula, which is apparently the most important queueing result. See, e.g., [6], p.60. It is well-known (and easily argued by Little's rule) that this value for the mean waiting time does not vary with the service regime as long as preemption (i.e., interrupting service during its execution) is not allowed and as long as non-anticipation (i.e., order of service is not based on prior knowledge of the actual service times) is assumed.

Two other well-known service regimes are that of Last Come First Served with Preemption Resume (LCFS-PR) and Egalitarian Processor Sharing (EPS). Under the former regime, last to arrive have preemptive priority over those who have arrived earlier. Customers might be preempted while in service, and when they return to service, it is resumed from the point where is was interrupted last. Under the EPS regime, the server splits its service capacity evenly among all those who are present in the system at any given time instant. This means that if n customers are present, they all receive a service of length $\Delta t/n$ during a period of length Δt (assuming Δt is short enough and no change in n occurs). It is well known that under these two schemes the mean time in the system for a customer whose service time equals x, is $x/(1-\rho)$. See, e.g., [6], p.63. In particular, the mean time in the system equals

$$\frac{\overline{x}}{1-\rho}. \tag{2}$$

Since we are concern in this article only with mean values, we next refer only to the LCFS-PR regime but whatever we derive is applicable to the EPS regime as well. In particular, we denote by $\mathrm{E}(W_{LCFS-PR})$ the common mean waiting time.

Which of the two regimes, FCFS or LCFS-PR, is better? From the question regarding the mean waiting time the answer is clear: Comparing (1) with (2), FCFS comes with a lower mean time in the system, namely $\frac{\lambda \overline{x^2}}{2(1-\rho)} + \overline{x} < \frac{\overline{x}}{1-\rho}$, if and only if $\mathrm{CoV}(X) < 1$, where $\mathrm{CoV}(X) = \sqrt{\overline{x^2} - \overline{x}^2}/\overline{x}$ (the ratio between the standard deviation of service time and its mean value). The opposite is the case where $\mathrm{CoV}(X) > 1$. They are equal when $\mathrm{CoV}(X) = 1$, which is for example the case where service times follow exponential distribution.[4]

Nevertheless, mean time in the system is not necessarily the single criterion to look at. FCFS looks fair and the norm in many cases. It also does not discriminate between customers based on their service requirement (which can be looked at a plus or a minus depending on the eye of the beholder). In particular, long jobs may need to hang around for a long time. Thus, FCFS can be the preferred discipline even when $\mathrm{CoV}(X) > 1$, i.e., when it comes with a higher mean waiting time. On the other hand, LCFS-PR and EPS have the theoretical advantage that the mean waiting time exists even in the case where the second moment of service does not exists (or, less formally, when $\overline{x^2} = \infty$).

The PRP qeueing regime that we suggest will be shown to be the 'middle of the road': The resulting mean time in the system lies between the corresponding means under the FCFS and LCFS-PR regimes. Thus, in the case where $\mathrm{CoV}(X) > 1$, by adopting the suggested scheme, one can do better in terms of reducing the mean sojourn time in comparison with the FCFS regime, without having to 'starve' the long jobs in the same scale as the LFCS-PR or the EPS regimes do.

Under the PRP scheme each arrival draws a random number which is uniformly distributed between zero and one. This number determines his preemptive priority level. We adopt the convention that the lower is the number drown, the higher the priority is. Specifically, one who draws a value of p is served before one who draws a value of q when $p < q$, possibly preempting the latter if found in service upon the arrival of the former. A

[4]Its easy to see that $\mathrm{CoV}(X) < 1$ ($\mathrm{CoV}(X) > 1$, respectively) if and only if $\overline{x^2}/2\overline{x} < \overline{x}$ ($\overline{x^2}/2\overline{x} < \overline{x}$, respectively). This means that the mean residual service time of a customer who is currently in service is smaller (larger, respectively) than the mean service time of a 'fresh' customer. For more on this concept see, e.g., [6], Chapter 2.

preempted customer resumes service when his turn (based, again, on his priority parameter) comes, from the point where it was interrupted last. Denote by $\mathrm{E}(W_{PRP})$ the resulting mean time in the system. We show below that $\mathrm{E}(W_{PRP})$ lies between $\mathrm{E}(W_{FCFS})$ and $\mathrm{E}(W_{LCFS-PR})$ (where, of course, the complete order is determined by comparing the service coefficient of variation of X to one). Note that had preemption not been allowed, the resulting mean waiting value would coincide with that of FCFS. Also note that since all continuous lotteries are monotone transformations of each other, the assumption of priorities determined by a uniform distribution is without loss of generality.

Remark. Note that in order to operate the PRP regime there is no need to maintain the order in which customers which are currently present had arrived (as is the case of the FCFS and the LCFS-PR regimes) or the order in which they receive service last (as in the case EPS regime when looked as the limit of the round-robin scheme).

Theorem 2.1 *The mean waiting time in the preemptive random priority (PRP) M/G/1 queue equals*

$$\mathrm{E}(W_{PRP}) = \frac{\rho + (1-\rho)\ln(1-\rho)}{(1-\rho)\rho^2}\frac{\lambda\overline{x^2}}{2} - \frac{\ln(1-\rho)}{\rho}\overline{x} \tag{3}$$

or, alternatively,

$$\mathrm{E}(W_{PRP}) = \frac{1}{1-\rho}\frac{\overline{x^2}}{2\overline{x}} - \frac{\ln(1-\rho)}{\rho}(\overline{x} - \frac{\overline{x^2}}{2\overline{x}}). \tag{4}$$

In particular, $\mathrm{E}(W_{PRP})$ *is bounded between* $\mathrm{E}(W_{FCFS})$ *and* $\mathrm{E}(W_{LCFS-PR})$. *Specifically, if* $CoV(X) < 1$ *then*

$$\mathrm{E}(W_{FCFS}) < \mathrm{E}(W_{RPR}) < \mathrm{E}(W_{LCFS}),$$

where the inequalities are reversed in the case where $CoV(X) > 1$.

See Figures 1 and 2 below for two examples.

Proof : Denote by P the random priority level that an arbitrary customer draws. Tag a customer and set his priority parameter to p. It means that a mass of size p gets preemptive priority over him. Then, by, e.g., [10], p.125, or [6], pp.76-77, his mean time in the system, equals

$$\mathrm{E}(W|P=p) = \frac{\lambda p}{2(1-p\rho)^2}\overline{x^2} + \frac{1}{1-p\rho}\overline{x}. \tag{5}$$

Clearly, $\mathrm{E}(W_{PRP}) = \mathrm{E}(\mathrm{E}(W|P))$. Our proof for (3) and (4) is completed by integrating the right-hand-side of (5) with respect to p from $p = 0$ through $p = 1$. The second part of the theorem is immediate knowing all three values and using a little bit of algebra. ■

Clearly, $\lim_{\rho \to 0} \mathrm{E}(W_{PRP}) = \overline{x}$ and $\lim_{\rho \to 1} \mathrm{E}(W_{PRP}) = \infty$. The following corollary says what is the relative performance of the PRP in comparison with the FCFS and the LCFS-PR regimes under heavy traffic. Its proof is straightforward and hence omitted.

Corollary 2.2

$$\lim_{\rho \to 1} \frac{\mathrm{E}(W_{PRP})}{\mathrm{E}(W_{LCFS-PR})} = \frac{\overline{x^2}}{2\overline{x}}$$

and

$$\lim_{\rho \to 1} \frac{\mathrm{E}(W_{PRP})}{\mathrm{E}(W_{FCFS})} = 1.$$

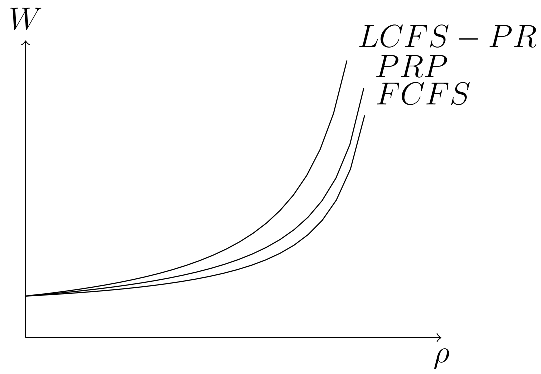

Figure 1: Mean waiting times under the three service regimes in the case $\bar{x} = 1$ and $CoV(X) = 0$ as a function of ρ.

3 Distribution of time in the system

Before moving on we like to introduce the concept of a *stand-by customer*. We define a stand-by customer as one who is singled out to receive service only when the system is otherwise empty. In particular, he is always preempted from service when someone else arrives. Using the terminology of the previous

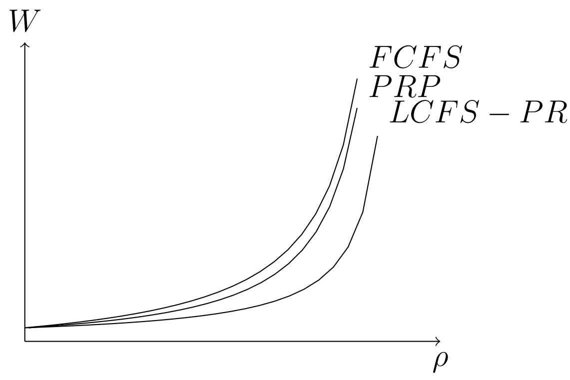

Figure 2: Mean waiting times under the three service regimes in the case $\bar{x} = 1$ and $CoV(X) = 2$ as a function of ρ.

section, he is the one, and only one, who has drawn a priority parameter 1. Likewise, from the point of view of the customers who possesses a priority parameter less than or equal to p, the customer with parameter p is a stand-by customer. Note that a stand-by customer does not inflict any extra waiting on any of the other customers. An economist would say here that his decision whether or not to join the queue does not come with any *externalities.*

Our interest in this section is in the LST of the time in the system under the PRP regime. Below we deal first with the M/M/1 case and then with the more general M/G/1 case. Admittedly, one can derive the former case from the latter but with a minimal cost in space we do the special case first while utilizing its special features and then we switch to the more general case.

3.1 The M/M/1 case

Let $B^*(s)$ denoted the LST of a busy period in an M/M/1 queue. It is well-known that if the arrival rate is λ and μ is the service rate (recall that $\mu^{-1} = \overline{x}$) then (assuming $\lambda < \mu$)

$$B^*(s) = \frac{\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}}{2\lambda}. \qquad (6)$$

See, e.g. [6], p.91.

Lemma 3.1 *The LST of the time in the system for a stand-by customer in an M/M/1 queue equals*

$$\frac{(1-\rho)B^*(s)}{1-\rho B^*(s)} \tag{7}$$

where $B^*(s)$ can be read from (6). In summary, it equals

$$\frac{2(\mu-\lambda)/\rho}{\lambda+\mu+s+\sqrt{(\lambda+\mu+s)^2-4\lambda\mu}} - \frac{1-\rho}{\rho} \tag{8}$$

Proof : A customer who sees upon his arrival $n \geq 0$ customers, has to stay in the system for time which is the sum of $n+1$ independent and identically distributed busy periods. The (conditional) LST is then $(B^*(s))^{n+1}$. The probability of seeing this number is $(1-\rho)\rho^n$, $n \geq 0$. See, e.g.,[6], p.120. Hence, the unconditional LTS equals

$$\sum_{n=0}^{\infty}(1-\rho)\rho^n(B^*(s))^{n+1} = \frac{(1-\rho)B^*(s)}{1-\rho B^*(s)}$$

as required. Deriving (8) can now done with the use of (6) and some algebra.
■

Corollary 3.2 *Denoted by W_{st} the time in the system for a stand-by customer. Then*

$$\mathrm{E}(W_{st}) = \frac{1}{(1-\rho)^2\mu} \tag{9}$$

and

$$Var(W_{st}) = \frac{1+2\rho}{(1-\rho)^4\mu^2} \tag{10}$$

[5]

Proof : A stand-by customer who finds $L \geq 0$ customers in the system has to wait $L+1$ busy periods prior to his departure. The mean length of a busy period equals $1/[(1-\rho)\mu]$ and $\mathrm{E}(L) = \rho/(1-\rho)$. By multiplying these two terms, one gets (9). Next we hint how we got $\mathrm{Var}(W_{st})$. In order to find the second moment of W_{st} one needs to take the second derivative of (7) and insert $s = 0$. It will be then useful to know that the first moment of the

[5] Formula (10) corrects an error in the corresponding formula which appears in the proceedings paper [7].

busy period equals $1/[(1-\rho)\mu]$ and its second moment equals $2/[(1-\rho)^3\mu^2]$. See, e.g., [6], p.92. From that, of course, one needs to subtract the square of the first moment of W_{st}. This moment, as we have already noted, equals $1/[(1-\rho)^2\mu]$. ■

As said, a customer who draws a priority parameter of p, his time in the system is as that of a stand-by customer in an M/M/1 but with an arrival rate of λp or, equivalently, with a traffic intensity of $p\rho$. Using (8), we conclude that the LST of his time in the system, denoted next by $T_p^*(s)$, equals

$$T_p^*(s) = \frac{2(\mu - \lambda p)}{\lambda p + \mu + s - \sqrt{(\lambda p + \mu + s)^2 - 4\lambda p\mu}} - \frac{1-\rho p}{\rho p} \tag{11}$$

Finally, the corresponding LST of the sojourn time of a random customer equals

$$\int_{p=0}^{1} T_p^*(s)\, dp.$$

The mean waiting time in case of the PRP regime is of course $1/[(1-\rho)\mu]$. We next state the variance of this time.

Theorem 3.3 *The variance of time in the system in the case of a PRP regime is denoted and equal to*

$$Var(W_{PRP}) = \frac{3+\rho^2}{3(1-\rho)^3\mu^2}.$$

In particular, the corresponding coefficient of variation equals

$$\sqrt{\frac{3+\rho^2}{3(1-\rho)}}.$$

Proof : Denote by U the random priority and recall that it is uniformly distributed in the unit interval. For simplicity in throughout this proof we replace W_{PRP} with W. We next derive Var(W) via the decomposition formula

$$\text{Var}(W) = \text{E}(\text{Var}(W|U)) + \text{Var}(\text{E}(W|U)).$$

Since a customer whose priority parameter equals U is in fact a stand-by customer in a system with traffic intensity ρU, we get from (10) that $\text{Var}(W|U) = \frac{1+2\rho U}{(1-\rho U)^4\mu^2}$. Hence,

$$\text{E}(\text{Var}(W|U)) = \int_{u=0}^{1} \frac{1+2\rho u}{(1-\rho u)^4\mu^2}\, du = \frac{1}{(1-\rho)^3\mu^2}.$$

As for the other term, our point of departure (due to (9)) is that $\mathrm{E}(W|U) = 1/[(1-\rho U)\mu]$. Hence,

$$\mathrm{Var}(\mathrm{E}(W|U)) = \mathrm{E}(\mathrm{E}^2(W|U)) - \mathrm{E}^2(\mathrm{E}(W|U)) = \mathrm{E}(\mathrm{E}^2(W|U)) - \mathrm{E}^2(W)$$

$$= \frac{1}{\mu^2}\int_{u=0}^{1} \frac{1}{(1-\rho u)^4}\,du - \frac{1}{(1-\rho)^2\mu^2}.$$

The rest is trivial algebra. ■

3.2 The M/G/1 case

It is well known that the LST of the time in the system for a customer in a FCFS M/G/1 equals

$$W^*(s) = (1-\rho)\frac{sG^*(s)}{\lambda G^*(s) + s - \lambda} \tag{12}$$

where $G^*(s)$ is the LST for a single service time. See, e.g., [6], p.86. Note that this is also the LST of the total workload which is held in the system upon arrival instants (as well as at any random instant). We denote by $B^*(s)$ the LST of a standard busy period. There is no explicit expression for this transform (the case of an M/M/1 considered is the previous section (see (6)) is an exception). Yet, it is known that $B^*(s)$ obeys the condition

$$B^*(s) = G^*(s + \lambda(1 - B^*(s))). \tag{13}$$

This identity can lead to the finding of all the moments of a busy period. See, e.g., [6], p.92, for the first two. They are $\overline{x}/(1-\rho)$ and $\overline{x^2}/(1-\rho)^3$, respectively. Consider now the time in the system of a stand-by customer. We next look for the LST of his time in the system.

Theorem 3.4 *Denote by $T^*(s)$ the LST of the time spend in the system by a stand-by customer in an M/G/1 queue. Then,*

$$T^*(s) = (1-\rho)\frac{(s + \lambda(1 - B^*(s)))B^*(s)}{s}.$$

We next give two proofs for this theorem.

Proof 1: It is possible to see that his time there coincides with a non-standard busy period, namely a busy period whose first service time is different than that of all others (whose distribution is $G(x)$). Moreover, this first service time is in fact the total work he finds in the system (inclusive is own) upon his arrival. Note that the LST of this 'special service' appears in (12). One can find in the literature the LST of a non-standard busy period. Specifically, denoting by $G_0(s)$ the LST of the special service, then the LST of the non-standard busy period equals

$$G_0(s + \lambda - \lambda B^*(s)).$$

See, e.g., [6], p.95. All needs to be done now, is to use $W^*(s)$ as given in (12) as $G_0(s)$. We can hence conclude, after some algebra (and with the help of (13)) that

$$\begin{aligned} T^*(s) &= W^*(s + \lambda - \lambda B^*(s)) \\ &= (1-\rho)\tfrac{(s+\lambda-\lambda B^*(s))G(s+\lambda-\lambda B^*(s))}{\lambda G^*(s+\lambda-\lambda B^*(s))+s-\lambda B^*(s)} \\ &= (1-\rho)\tfrac{(s+\lambda(1-B^*(s)))B^*(s)}{s}, \end{aligned} \tag{14}$$

as required. ∎

Proof 2: In the case where he arrives to an empty system, a probability $1-\rho$ event, his conditional LST equals $B^*(s)$. Otherwise, a probability ρ event, he needs to wait for the residual of the running busy period. This comes with a LST of $(1 - B^*(s))/(s\mathrm{E}(B))$ where $\mathrm{E}(B)$, the mean busy period, which equals $\overline{x}/(1-\rho)$. See, e.g., [6], p.30. This needs to be multiplied by $B^*(s)$ due to the busy period which initiates as soon as he enters service for the first and is concluded upon his departure. In summary,

$$T^*(s) = (1-\rho)B^*(s) + \rho\frac{1 - B^*(s)}{s\overline{x}}(1-\rho)B^*(s). \tag{15}$$

The rest is algebra. ∎

Theorem 3.5 *The mean sojourn time of a stand-by customer in an M/G/1 queue equals*

$$\frac{\lambda\overline{x^2}}{2(1-\rho)^2} + \frac{\overline{x}}{1-\rho}. \tag{16}$$

We next suggest three proofs for this theorem.

Proof 1: Take the derivative with respect to s in (15) and insert $s = 0$. The call for the L'Hospital rule will be required. In particular, the second moment of the busy period will be required. It equals $\overline{x^2}/(1-\rho)^3$ (see, e.g. [6], pp.92). We omit any further details. ∎

Proof 2: Upon arrival, the amount of work found by the stand-by customer (or by any body else) and inclusive of his own, equals

$$\frac{\lambda\overline{x^2}}{2(1-\rho)}+\overline{x}. \tag{17}$$

This needs to be divided by $1-\rho$ in order to find the mean time until the system is emptied for the first time (see, e.g. [6], p.63), an instant of time which coincides with the instant in which the stand-by customer departs. ■

Proof 3: A third way is to look into the model as one with preemptive priority with two classes. All are in the high priority class, while a single customer (who represents a zero arrival rate class) is inferior. Both classes of course share the same first two moments $\overline{x}$ and $\overline{x^2}$. See, e.g.,[10], p.125 or [6], p.76-77, which gives the mean sojourn time for customers of both classes. For the inferior class who has an arrival rate of zero, this means coincides with (16). ■

Remark. From [9], we learn that the expected marginal externalities that a customer inflicts on others in case of a FCFS regime, equals

$$\frac{\lambda\overline{x^2}}{2(1-\rho)^2}.$$

To this we need to add his own mean sojourn time, which is stated in (1), in order to find out the total social costs inflicted by an arrival. As we can see, we do not get the same value that we got for the mean time in the system for a stand-by customer. Yet, the two coincide in case of an exponential service time (something which can be checked with minimal algebra). The reason behind that is that in comparing systems, one with an extra customer and the other without, under the same arrival and service processes, one gets always one more customer in the former case from one's arrival until the system is empty for the first time, only if one assumes exponential service times. Hence, the mean time for a stand-by customer and mean added social costs coincide in case of exponential service. This is not the case under a general service distribution.

For a customer who draws a priority level of p, the LST of his time in the system is as above where λ is replaced by λp. Specifically, denote this LST by $T_p^*(s)$ and conclude from (14) that

$$T_p^*(s)=(1-p\rho)\frac{(s+\lambda p(1-B_p^*(s)))B_p^*(s)}{s}.$$

where $B_p^*(s)$ is the LST of the busy period but with an arrival rate of λp, rather than λ.[6] Likewise, when we look for the mean value, we need to replace in (16) λ with λp and ρ with $p\rho$. In particular, the corresponding mean equals

$$\frac{p\lambda\overline{x^2}}{2(1-p\rho)^2} + \frac{\overline{x}}{1-p\rho} \tag{18}$$

The following theorem is now immediate.

Theorem 3.6 *Denote by $T^*_{PRP}(s)$ the LST of the sojourn time of a customer in a PRP M/G/1 queue. Then,*

$$T^*_{PRP}(s) = \int_{p=0}^{1} (1-p\rho)\frac{(s+\lambda p(1-B_p^*(s)))B_p^*(s)}{s}\,dp.$$

As for the mean value, we can by-pass the need to have in hand first the LST. Firstly, it was derived independently in Theorem 2.1. Secondly, we use (18) and derive

$$\mathrm{E}(W_{PRP}) = \int_{p=0}^{1} \left(\frac{\lambda p\overline{x^2}}{2(1-p\rho)^2} + \frac{\overline{x}}{1-p\rho}\right) dp.$$

Of course, one should get the same result as in (3) and in (4).

4 Equilibrium behavior and server utilization

Assume now that customers gain a value of R due to service completion and it costs each one of them C per unit of time in the system (service inclusive). Thus, $R-CW$ is the mean net return from joining when W denotes the mean time in the system. Without loss of generality, we assume that not joining comes with a zero reward (otherwise, one would need to shift R accordingly). It would then makes sense to assume that one joins if and only if one's net gain from joining is positive. Consider W now as a function of the arrival rate and denote it hence by $W(y)$ when y is the arrival rate. Assume now that λ, which as of now will be looked at as the *potential* arrival rate, is such that $R - CW(\lambda) < 0$. In other words, if all join, one is better off not joining. Assume also that $R > C\overline{x}$, namely the reward is larger than

[6] See (11) for the case of M/M/1.

the cost of time due just to the time in service. Then, one is better off joining when no-body else does that. We have reached a circular reasoning. The Nash equilibrium concept from non-cooperative game theory deals with such cases. In particular, it implies that each customer should join with a probability of p_e where p_e solves $R - CW(\lambda p_e) = 0$. Indeed, when all join with a probability of p_e, one is indifferent between joining and not, and hence one is willing to randomize between the two options with any probability, p_e inclusive. For more on this concept in the context of queues, see [1] and [5].

Assuming that customers behave in accordance with such an equilibrium behavior is somewhat sad news: Those who join, as well as those who do not join, end up with nothing. Nothing is also the social gain, or the consumer surplus, here. Moreover, changing the function W (for example, by changing the queue regime or by changing the service rate), does not lead to any individual or social gain. Indeed, for commuters who use a road which is usually jammed, adding another lane will not help: Once this is done, more commuters will use the road, leading to the same (slow) traffic speed.

The previous paragraph implies that, from the customers point of view, it indeed does not matter which queueing regime is used. But this is certainly not the case from the point of view of the server (or the one who owns the service facility). This is the case since the server utilization in equilibrium, which equals $\lambda p_e \overline{x}$, may vary with the regime due to the simple reason that p_e varies with it.

Theorem 4.1 *Denote by p_e^{FCFS} the Nash equilibrium joining probability under the queueing regime is FCFS. Define p_e^{PRP} and $p_e^{LCFS-PR}$ in a similar fashion. Then, if $CoV(X) < 1$,*

$$p_e^{FCFS} < p_e^{PRP} < p_e^{LCFS-PR}.$$

The inequalities are reversed when $CoV(X) > 1$.

The proof of this theorem is now straightforward and we omit the details. An example for a consequence from this theorem is that if $CoV(X) > 1$ and the current norm is to use FCFS, switching to PRP will decrease the server utilization (and hence the server might be useful for some other functions) without effecting the social benefit. Of course, one might imagine cases in which one likes to increase the server's utilization. Hence, such a switch will be recommended in the case where $\mathrm{CoV}(X) < 1$.

5 The case of non-preemptive priority and future research

A possible variation of the regime stated above (still in the M/G/1 context) is that priority will be granted but without preemption. The mean time in the queue (service inclusive) of one whose drawn priority parameter equals U is

$$\frac{W_0}{(1-U\rho)^2}.$$

where $W_0 = \lambda\overline{x^2}/2$. See, e.g. [6]. p.60. The overall mean time in the system is as in the case of FCFS, namely it equals $W_0/(1-\rho)$ since this moment does not vary with the order in which customers enter service (as long as it is not done in a service dependent way). This is in opposed to the case with preemption whose overall performance changes due to preemption. Indeed, all work-conserving, non-anticipating and non-preempting regimes share the same mean queueing time. Thus, this regime is not recommended if one likes to move away from the standard FCFS. The focus may now be turned into higher moments or in fact the distribution at large. For the LST we can consult results on the LST of the queueing time distributions of various non-preemptive priority classes. In particular, from [10], p.121, we can get that the LST of the queueing time of a customer whose priority parameter equals p, denoted by $W_p^*(s)$, equals

$$\frac{(1-\rho)(s+\lambda p-\lambda p B_p^*(s))+\lambda(1-p)(1-G^*(\lambda+\lambda p-B_p^*(s)))}{s}.$$

The LST we are after equals then to

$$\int_{p=0}^{1} W_p^*(s)\,dp.$$

Acknowledgement

Comments made by Yoav Kerner and Binyamin Oz are highly appreciated. The research was supported by Israel Science Foundation grant no. 1319/11.

References

[1] Edleson, N.M. and D.K. Hildebrand (1975), "Congestion tolls for Poisson queueing process," *Econometrica*, **43**, 81-92.

[2] Glazer, A. and R. Hassin (1986), "Stable prioirity purchasing in queues," *Operations Reserch Letters,* **4**, 285-288.

[3] Harchol-Balter, M. (2013), *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, Cambridge University Press, Cambridge.

[4] Hassin, R. (1995), "Decentralized regulation of a queue," *Management Science,* **41**, 163-173.

[5] Hassin, R. and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behaviour in Queueing Systems,* Kluwer.

[6] Haviv, M. (2013), *Queueus - A Course in Queueing Theorey*, Springer.

[7] Haviv, M. (2014), "The M/G/1 queueing model with preemptive random priorities," *Proceedings of Valuetools 2014*, Bratislava 2014.

[8] Haviv, M. and B. Oz (2014), "Self-regulation of a queue," submitted.

[9] Haviv, M. and Y. Ritov (1998), "Externalities, tangible externalities and queue disciplines," *Management Science*, **44**, 850-858.

[10] Kleinrock, L. (1976) *Queueing systems, Volume 2: Computer applications*, John Wiley and Sons, New York.