2005 American Control Conference
June 8-10, 2005. Portland, OR, USA

Average Rate in a M/M/1 Processor-Sharing Queue

Na Chen and Scott Jordan

Abstract— **We investigate the average rate per job in an open M/M/1 processor-sharing queue. We introduce three definitions of average rate as seen by the system and by each job, and present expressions for each in terms of the system rate and load. We compare the three measures and prove that they are strictly ordered over all loads. We next consider the system rate required to achieve a minimum average rate per job, and prove that it is increasing and convex. Finally, we consider the system rate required to achieve a minimum tail probability on the average rate per job, and present expressions illustrating when the system rate is constrained by the required tail probability.**

I. Introduction

Processor-sharing (PS) queues have long been of interest in a wide variety of operations research environments, including telecommunications. We are particularly motivated by data telecommunication systems, in which each user represents a file to transmit and the queue service rate represents the total transmission rate of the system. If the total rate of the system is shared equally between all active file transmissions, then the system can be modeled by a processor-sharing queue.

The most common data telecommunication performance measures involve the *transmission rate per user*. From the system's perspective, the total transmission rate is split among the active users, and therefore the instantaneous rate per user changes whenever a user arrives or departs. From the user's perspective, the average transmission rate is defined as the file size divided by the time required to transmit the file. In the context of a processor-sharing queue, these transmission rates per job are therefore related to the stationary distribution of the queue or to the ratio of a user's service time requirement to the user's sojourn time. There is a long literature on processor-sharing systems, including characterizations of the stationary distribution of the queue and the distribution of users' sojourn times. However, we have found no literature on the transmission rates per job.

For open M/M/1-PS systems, Coffman et. al. [1] derived the Laplace transform of the waiting time distribution of a tagged user, conditioned on the required service time and the number in the system upon the tagged arrival. From this result, they obtained the first two moments of the conditioned equilibrium waiting time. By removing the conditioning and inverting this Laplace transform, Morrison [2] obtained an integral representation for the complementary distribution of the sojourn time, which was refined by Guillemin and Boyer [3] via spectral theory to obtain the distribution of the sojourn time of a user conditioned on the number of users in the system at its arrival. The moments of this distribution were further studied by Sengupta and Jagerman [4].

For closed M/M/1-PS systems, Morrison obtained the asymptotic approximation to the equilibrium distribution of the waiting time [5], as well as the distribution of the response time conditioned on the required service time in the very heavy-usage case [6], using perturbation analysis on the generating functions of the sojourn time. Similar perturbation techniques were used by Knessl [7] to construct an asymptotic approximation to the sojourn time distribution in a large heavy loaded system. Gaver and Jocobs [8] investigated the multiclass processing-sharing queues using a heavy traffic diffusion approximation. Switching time were considered by Bersani [9] and Barbagallo [10].

There is also a literature on other types of processor-sharing queues. The steady-state behavior of two M/M/1 parallel PS queues under the head-of-the-line PS discipline is investigated in [11] and [12]. GI/M/1-PS systems were studied in [13]-[17], M/G/1-PS systems in [18]-[19], MAP/M/1-PS systems in [20].

Finally, there is a literature on *slowdown* in queues, defined as the ratio of the sojourn time to the job size (c.f. [21]). *Mean slowdown* is often used as a measure of system performance as opposed to the more traditional mean sojourn time. Under processor-sharing, all jobs have the same mean slowdown; hence the processor is fairly shared among all jobs in the system.

Our focus, however, is on the transmission rate per job. We present three definitions of average rate per job. The first two are from the system's perspective, and are related to the stationary distribution of the queue. The third is from the user's perspective, and is related to the ratio of a user's service time requirement to the user's sojourn time. Although the literature on the stationary distribution, sojourn time, and slowdown ratio for processor-sharing queues forms a foundation for our work, it does not directly address these performance measures.

The remainder of this paper is organized as follows. In section II, we present the three definitions of average rate per job. We derive expressions for each in terms of the system rate and load, and prove that they are strictly ordered over all loads. In section III, we consider two types of performance bounds, on average rate and on the probability of meeting or exceeding a specified rate. We prove that the system rate required to achieve a minimum average rate per job is increasing and convex. We also present expressions illustrating when the system rate is constrained

This material is based upon work supported by the National Science Foundation under Grant No. ANI-0137103 and by DARPA under Grant No. N66001-00-8935.

Both authors are with the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, 92697-2625 {nac,sjordan}@uci.edu

by the required tail probability.

II. Queueing model

We are motivated by users sharing a data channel of total transmission rate R (bits/sec). Suppose that users arrive as a Poisson process with rate λ (jobs/sec). Suppose that each user transmits a file and then departs from the system, and that file sizes are independent and identically distributed and drawn from an Exponential distribution with mean F (bits). Suppose that when there are $N \geq 1$ jobs in the system, each user transmits at an instantaneous rate of R/N.

The system is therefore equivalent to a M/M/1-PS queue with arrival rate λ and service rate μ (jobs/sec), where $\mu = R/F$. Denote the load on the queue by $\rho \equiv \lambda/\mu$, and assume that $\rho < 1$. It follows that the stationary distribution for the number of users in the system is given by:

$$\pi_n \equiv Pr(N = n) = (1-\rho)\rho^n, n = 0, 1, 2, \cdots$$

We are interested in the transmission rate of users when the system is not empty, and therefore we will need to know the distribution of the number of jobs in the queue during busy cycles, given by:

$$\begin{aligned} p_n &\equiv Pr(N = n | n \geq 1) \\ &= \frac{\pi_n}{1-\pi_0} = (1-\rho)\rho^{n-1}, n = 1, 2, \cdots \end{aligned} \quad (1)$$

A. Three definitions of average rate

We now introduce three definitions of *average rate* in the system. We start with the average rate per user as seen by the system, conditioned on at least one user in the system:

Definition 1:

$$m_{time} \equiv \sum_{n=1}^{\infty} \frac{R}{n} p_n \quad (2)$$

A closed form expression for m_{time} in terms of either R and ρ or R and λF can be found by substituting (1) into (2) and expressing the sum as a logarithm:

$$\begin{aligned} m_{time} &= \frac{R(1-\rho)}{\rho} \ln(1/(1-\rho)) \\ &= \frac{R(R-\lambda F)}{\lambda F} \ln(R/(R-\lambda F)) \end{aligned} \quad (3)$$

A second definition of average rate can be created by weighting the instantaneous rate per user by the number of users in the system. Instead of using the stationary distribution during a busy cycle $\{p_n\}$, we weight the distribution by the number of jobs n, namely we replace (1) by $p'_n \equiv Cn\rho^n$. The constant C can be found by evaluating $\sum_{n=1}^{\infty} p'_n = 1$, resulting in:

$$p'_n = n(1-\rho)^2\rho^{n-1}, n = 1, 2, \cdots \quad (4)$$

Our second definition of average rate is thus:

Definition 2:

$$m_{weighted} = \sum_{n=1}^{\infty} \frac{R}{n} p'_n \quad (5)$$

Substituting (4) into (5) yields a closed form expression in terms of either R and ρ or R and λF:

$$m_{weighted} = R(1-\rho) = R - \lambda F \quad (6)$$

Finally, a third definition of average rate is from the perspective of the users. Let F_i represent the file size of user i. Let T_i represent user i's sojourn time, the time user i spends in the system from arrival until completion of service. Let R_i represent user i's average rate (over its sojourn time). It follows that $R_i = F_i/T_i$.

The third definition of average rate is R_i averaged over all users:

Definition 3:

$$m_{jobs} = E_i R_i \quad (7)$$

Since the arrivals form a Poisson process, an arrival will find the number of users in the system (excluding itself) to reflect the stationary distribution $\{\pi_n\}$. However, over a user's sojourn time, the expected average rate depends not only upon the stationary distribution $\{\pi_n\}$, but also upon transients. The problem is that a user's sojourn time *also* depends on future arrivals and departures, due to the processor-sharing service discipline.

The key is to condition on customer's i's file size, or equivalently upon its service time. Denote the average rate for jobs with file size $R\tau$ (and hence service time τ) by:

$$\begin{aligned} Y(\tau) &= E_i(R_i | F_i = R\tau) = E_i(F_i/T_i | F_i = R\tau) \\ &= R\tau E_i(1/T_i | F_i = R\tau) \end{aligned} \quad (8)$$

It follows that $m_{jobs} = E_\tau Y(\tau)$.

The Laplace transform of sojourn time conditioned on a job's service time is given in [1]:

$$E_i(e^{-sT_i} | F_i = R\tau) = c_1(\lambda, \mu, s, \tau) \quad (9)$$

where

$$c_1(\lambda, \mu, s, \tau) = \frac{(1-\rho)(1-\rho r^2)e^{-\lambda(1-r)\tau}e^{-s\tau}}{(1-\rho r)^2 - \rho(1-r)^2 e^{-\mu\tau(1-\rho r^2)/r}}$$

and

$$r = \frac{\lambda + \mu + s - [(\lambda+\mu+s)^2 - 4\lambda\mu]^{\frac{1}{2}}}{2\lambda} \quad (10)$$

This result can be used to find an expression for $E_i(1/T_i | F_i = R\tau)$ by integrating the conditional Laplace transform:

$$\begin{aligned} &\int_0^\infty E_{T_i}(e^{-sT_i} | F_i = R\tau)\, ds \\ &= E_{T_i}\left(\int_0^\infty e^{-sT_i}\, ds \middle| F_i = R\tau \right) \\ &= E_{T_i}\left(\frac{1}{T_i} \middle| F_i = R\tau \right) \end{aligned} \quad (11)$$

The average rate for jobs with service time τ is thus given by substituting (9) and (11) into (8):

$$\begin{aligned} Y(\tau) &= R\tau \int_0^\infty E_i(e^{-sT_i} | F_i = R\tau)\, ds \\ &= R\tau \int_0^\infty c_1(\lambda, \mu, s, \tau)\, ds \end{aligned}$$

Furthermore, since F_i is Exponentially distributed with mean F, the average rate as seen by users is:

$$\begin{aligned} m_{jobs} &= E_\tau Y(\tau) \\ &= \int_0^\infty \frac{R\tau}{F} e^{-\frac{R\tau}{F}} \left(\int_0^\infty c_1(\lambda,\mu,s,\tau)\, ds \right) d(R\tau) \\ &= \frac{R^2}{F} \int_0^\infty \tau e^{-\frac{R\tau}{F}} \left(\int_0^\infty c_1(\lambda,\mu,s,\tau)\, ds \right) d\tau \end{aligned}$$

where r is given by (10).

In order to write m_{jobs} solely in terms of either R and ρ, we can solve (10) for s to obtain $s = u\mu$ where:

$$u = \frac{(1-r)(1-\rho r)}{r} \tag{12}$$

Substituting R/F by μ and using (12), we can represent m_{jobs} as:

$$m_{jobs} = R(1-\rho) \int_0^\infty \mu\tau e^{-\mu\tau} \left(\int_0^\infty c_2(\lambda,\mu,s,\tau)\, ds \right) d\tau \tag{13}$$

where

$$c_2(\lambda,\mu,s,\tau) = \frac{(1-\rho r^2) e^{-\frac{1-r}{r}\mu\tau}}{(1-\rho r)^2 - \rho(1-r)^2 e^{-\frac{(1-\rho r^2)}{r}\mu\tau}}$$

Define $v = \mu\tau$. Using a variable substitution from $\{s,\tau\}$ to $\{u,v\}$, we obtain:

$$m_{jobs} = R(1-\rho) \int_0^\infty v e^{-v} \left(\int_0^\infty c_3(\rho,u,v)\, du \right) dv \tag{14}$$

where

$$c_3(\rho,u,v) = \frac{(1-\rho r^2) e^{-\frac{1-r}{r}v}}{(1-\rho r)^2 - \rho(1-r)^2 e^{-\frac{(1-\rho r^2)}{r}v}}$$

and

$$r = \frac{1+u+\rho - [(1+u+\rho)^2 - 4\rho]^{\frac{1}{2}}}{2\rho}$$

B. Comparisons of average rates

In this subsection, we compare the three definitions of average rate presented above. The main result is given in the following theorem:

Theorem 1: $m_{time} > m_{jobs} > m_{weighted},\ \forall\ 0 < \rho < 1.$

Proof: We start by establishing that $0 < r < 1$. Since $4\lambda\mu > 0$, it follows from (10) that $r > 0$. To establish that $r < 1$, multiply both the numerator and denominator in (10) by $\lambda+\mu+s+[(\lambda+\mu+s)^2 - 4\lambda\mu]^{1/2}$:

$$\begin{aligned} r &= \frac{2\mu}{\lambda+\mu+s+[(\lambda+\mu+s)^2 - 4\lambda\mu]^{\frac{1}{2}}} \\ &< \frac{2\mu}{\lambda+\mu+s+[(\mu-\lambda+s)^2]^{\frac{1}{2}}} \\ &= \frac{2\mu}{2\mu+2s} < 1 \end{aligned} \tag{15}$$

where the first inequality follows from $(\lambda+\mu+s)^2 - 4\lambda\mu > (\mu-\lambda+s)^2$.

We now compare m_{time} as represented in (16) with m_{jobs} represented in (13). We start by expressing the term $\ln(1/(1-\rho))$ in (16) as a double integral with respect to s and τ:

$$\begin{aligned} & \ln(1/(1-\rho)) \\ &= \ln(1/(1-\rho)) \int_0^\infty \mu e^{-\mu\tau}\, d\tau \\ &= \int_0^\infty \mu e^{-\mu\tau} \left[\ln(1-\rho e^{-s\tau}) \right]_{s=0}^{s=\infty} d\tau \\ &= \int_0^\infty \mu\rho\tau e^{-\mu\tau} \left(\int_0^\infty \frac{e^{-s\tau}}{1-\rho e^{-s\tau}}\, ds \right) d\tau \end{aligned}$$

Then m_{time} in (3) can be written in a similar form as m_{jobs} in (14):

$$m_{time} = R(1-\rho) \int_0^\infty \mu\tau e^{-\mu\tau} \left(\int_0^\infty c_4(\lambda,\mu,s,\tau)\, ds \right) d\tau \tag{16}$$

where

$$c_4(\lambda,\mu,s,\tau) = \frac{e^{-s\tau}}{1-\rho e^{-s\tau}}$$

Comparing (16) with (13), it follows that a sufficient condition for $m_{time} > m_{jobs}$ to hold is:

$$c_4(\lambda,\mu,s,\tau) > c_2(\lambda,\mu,s,\tau),\ \forall\ \{s,\tau,\lambda\} > 0,\ 0 < \rho < 1.$$

Substitute s by $\mu(1-r)(1-\rho r)/r$ into this expression and and simplifying yields the equivalent sufficient condition:

$$\begin{aligned} & (1-\rho r)^2 e^{\frac{(1-\rho r^2)}{r}\mu\tau} \\ > \ & \rho(1-r)^2 + (1-\rho r^2) \left[e^{\frac{(1-\rho r)}{r}\mu\tau} - \rho e^{(1-\rho r)\mu\tau} \right] \end{aligned}$$

Finally, using the Maclaurin series of e^x and simplifying, if follows that $m_{time} > m_{jobs}$ if:

$$\sum_{n=3}^{\infty} \frac{(\mu\tau)^n}{n!} H_n(\rho,r) > 0$$

where

$$\begin{aligned} H_n(\rho,r) &= (1-\rho r)^2 \frac{(1-\rho r^2)^n}{r^n} \\ &\quad -(1-\rho r^2) \left[\frac{(1-\rho r)^n}{r^n} - \rho\,(1-\rho r)^n \right] \end{aligned}$$

We will show that the sum is positive by showing that each term is positive, using mathematical induction method.

$$H_3(\rho,r) = \frac{\rho(1-\rho r)^2(1-\rho r^2)(1-r)^2}{r^2}$$

Since $0 < \rho < 1$ and $0 < r < 1$, $H_3(\rho,r) > 0$ establishing the base case. For the induction case, assume that $H_k(\rho,r) > 0$ for some $k > 3$.

$$H_{k+1}(\rho,r) > \frac{\rho(1-\rho r)^k(1-\rho r^2)(1-r)}{r} \left(\frac{1}{r^{k-1}} - 1 \right)$$

Since $r < 1$, if follows that $H_{k+1}(\rho,r) > 0$. Consequently $H_n(\rho,r) > 0$, for $n \geq 3$, and hence $m_{time} > m_{jobs}$.

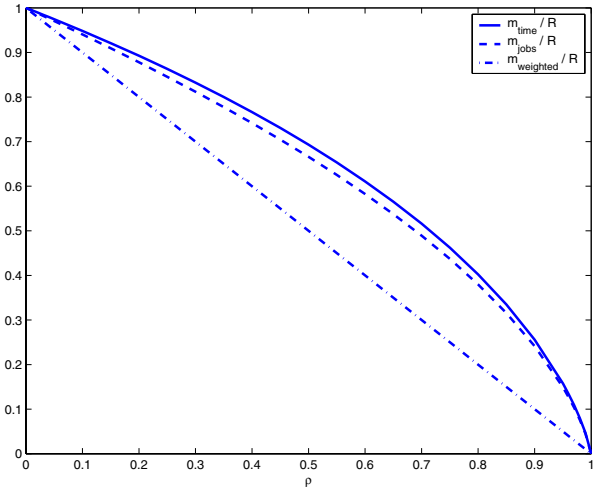

Fig. 1. comparisons of average rates

We now prove the second part of the theorem, that $m_{jobs} > m_{weighted}$. From (6) and (14), we have:

$$\frac{m_{jobs}}{m_{weighted}} = \int_0^\infty ve^{-v}\left(\int_0^\infty c_3(\rho,u,v)\,du\right)dv$$

Define $c_5(\rho,r,v) \equiv c_3(\rho,u,v)$. Using a variable substitution from u to r, we obtain:

$$\frac{m_{jobs}}{m_{weighted}} = \int_0^\infty ve^{-v}\left(\int_0^1 c_5(\rho,r,v)\frac{1-\rho r^2}{r^2}\,dr\right)dv$$

We can bound the integrand for all $0<\rho<1$, $0<r<1$ and $0<v<\infty$:

$$\begin{aligned} & ve^{-v}c_5(\rho,r,v)\frac{1-\rho r^2}{r^2} \\ = & \frac{(1-\rho r^2)^2}{(1-\rho r)^2-\rho(1-r)^2e^{-\frac{(1-\rho r^2)}{r}v}}\frac{v}{r^2}e^{-\frac{v}{r}} \\ \geq & \frac{(1-\rho r^2)^2}{(1-\rho r)^2}\frac{v}{r^2}e^{-\frac{v}{r}} > \frac{v}{r^2}e^{-\frac{v}{r}} \end{aligned}$$

It follows that

$$\frac{m_{jobs}}{m_{weighted}} > \int_0^\infty\left(\int_0^1 \frac{v}{r^2}e^{-\frac{v}{r}}\,dr\right)dv = 1$$

Therefore $m_{jobs} > m_{weighted}$. ■

The three average rates (normalized by R) are plotted versus the load ρ in figure 1. We observe that the difference between m_{time} and m_{jobs} is relatively small, and that the difference between m_{time} or m_{jobs} and $m_{weighted}$ is concave with respect to ρ.

C. Tail probability

One additional performance measure we are interested in is the tail probability of rate. Specifically, denote the instantaneous rate per job at time t by X_t. From the point of view of the system, under the stationary distribution conditioned on at least one job in the system (1), the distribution of X_t is given by:

$$P\left(X_t = \frac{R}{n}\right) = p_n = \left(1-\frac{\lambda F}{R}\right)\left(\frac{\lambda F}{R}\right)^{n-1},\ n \geq 1$$

Therefore the probability of receiving an instantaneous rate of x or better is given by:

$$P(X \geq x) = \sum_{n=1}^{k} P\left(X = \frac{R}{n}\right) = 1 - \left(\frac{\lambda F}{R}\right)^k$$

where $k = \lfloor R/x \rfloor \geq 1$.

We will consider the interpolated function:

$$G(x,R) \equiv 1 - \left(\frac{\lambda F}{R}\right)^{R/x} \tag{17}$$

which is a continuous approximation to $P(X \geq x)$ defined on $0 < x \leq R$.

III. Marginal bandwidths

In this section, we consider the ability of the system to provide two types of performance bounds: $m_{time} \geq m$ and $G(x,R) \geq p$. We consider them separately in the next two subsections. For purposes of discussion, we assume that the user arrival rate λ and the average file size F are fixed, but that the total system transmission rate R can be chosen by an appropriate investment into the system. We further assume that in all cases $R > \lambda F$, so that $\rho < 1$.

A. Bound on mean rate

We start by examining the derivative of m_{time} with respect to R, which follows from (3):

$$\frac{\partial m_{time}}{\partial R} = \left(\frac{2R}{\lambda F}-1\right)\ln\left(\frac{R}{R-\lambda F}\right) - 1$$

It can be easily shown that m_{time} increases monotonically with R:

Theorem 2: $\frac{\partial m_{time}}{\partial R} > 0\ \forall\ R$.

Proof: Expanding $\ln(\frac{R}{R-\lambda F})$ using a Maclaurin series,

$$\begin{aligned} \frac{\partial m_{time}}{\partial R} &= \left(\frac{2R}{\lambda F}-1\right)\sum_{n=1}^{\infty}\frac{(\lambda F/R)^n}{n} - 1 \\ &= 1 - \frac{\lambda F}{R} + \left(\frac{2R}{\lambda F}-1\right)\sum_{n=2}^{\infty}\frac{(\lambda F/R)^n}{n} > 0 \end{aligned}$$

■

From (3), in order to satisfy $m_{time} \geq m$, the rate $R \geq R_{min}$, where R_{min} is determined by the fixed point equation:

$$m = \frac{R_{min}(R_{min}-\lambda F)}{\lambda F}\ln\left(\frac{R_{min}}{R_{min}-\lambda F}\right) \tag{18}$$

It follows that the marginal bandwidth required for an increase in the average rate m_{time} is given by:

$$\frac{\partial R}{\partial m_{time}} = \left(\left(\frac{2R}{\lambda F}-1\right)\ln\left(\frac{R}{R-\lambda F}\right) - 1\right)^{-1}$$

A stronger characterization of R_{min} versus m is described in the following theorem:

Theorem 3: R_{min} is a monotonically increasing and convex function of m, and $R_{min} - m$ monotonically decreases with m from λF to $\lambda F/2$.

Proof: From Theorem 2, $\frac{\partial m}{\partial R_{min}} > 0$. Thus R_{min} monotonically increases from λF to infinity as m increases from 0 to infinity. Consider the second derivative, $\frac{\partial^2 m}{\partial R_{min}^2}$:

$$\frac{\partial^2 m}{\partial R_{min}^2} = \frac{\rho_{max}(\rho_{max}-2) - 2(1-\rho_{max})\ln(1-\rho_{max})}{R_{min}\rho_{max}(1-\rho_{max})}$$

where $\rho_{max} = \lambda F/R_{min}$.

The denominator is positive for $0 < \rho_{max} < 1$. Denote the numerator by $f_1(\rho_{max})$; it is negative for $0 < \rho_{max} < 1$ since

$$\lim_{\rho_{max}\to 0} f_1(\rho_{max}) = 0$$

and

$$f_1'(\rho_{max}) = 2\rho_{max} + 2\ln(1-\rho_{max}) < 0, \ \forall\ 0 < \rho_{max} < 1$$

Hence $\frac{\partial^2 m}{\partial R_{min}^2} < 0$ and thus R_{min} is a convex function of m.

To prove the variation of $R_{min} - m$ with m, expand $\ln(1 - \lambda F/R_{min})$ by its Maclaurin series, substituting in (18), and simplifying yields:

$$R_{min} - m = \left(\frac{1}{2} + f_2(\rho_{max})\right)\lambda F$$

where

$$f_2(\rho_{max}) = \sum_{n=1}^{\infty} \frac{\rho_{max}^n}{(n+1)(n+2)}$$

Now $0 < f_2(\rho_{max}) < 1/2$ for $0 < \rho_{max} < 1$. As m increases from 0 to infinity, we have already noted that R_{min} increases monotonically from λF to infinity. It follows that ρ_{max} decreases monotonically from 1 to 0, and therefore that $f_2(\rho_{max})$ decreases monotonically from $1/2$ to 0. This establishes that $R_{min} - m$ decreases monotonically from λF to $\lambda F/2$. ■

R_{min} versus m is shown in figure 2, along with its asymptote $R_{min} - m = \lambda F/2$. The asymptote can be thought of as a limit as the load approaches zero. In the limit, during a busy cycle there is one job in the system with probability $1 - \rho_{max}$, and two jobs with probability ρ_{max}. It follows that

$$m_{time} = \left(1 - \frac{\lambda F}{R_{min}}\right) R_{min} + \frac{\lambda F}{R_{min}} \frac{R_{min}}{2} = R_{min} - \frac{\lambda F}{2}$$

B. Bound on tail probability

We turn next to examining the bandwidth required for a bound on the tail probability of rate. We use the continuous interpolation $G(x, R)$ of $P(X \geq x)$, as defined in (17). We start by examining the derivative of $G(x, R)$ with respect to R, which can be shown to be:

$$\frac{\partial G(x,R)}{\partial R} = \left(\frac{\lambda F}{R}\right)^{R/x} \frac{1 + \ln(R/\lambda F)}{x}, R \geq x, R > \lambda F$$

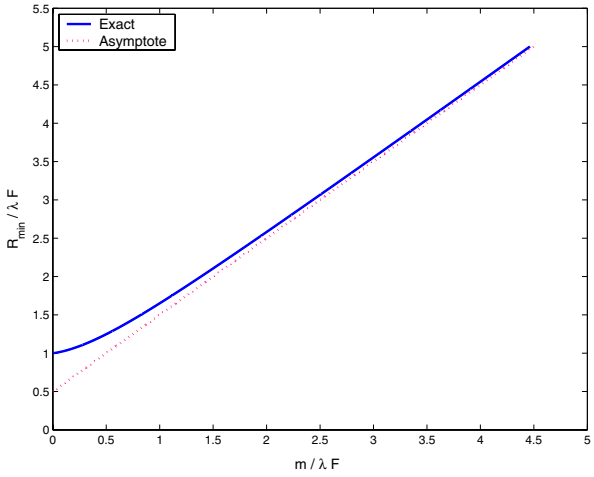

Fig. 2. R_{min} versus m

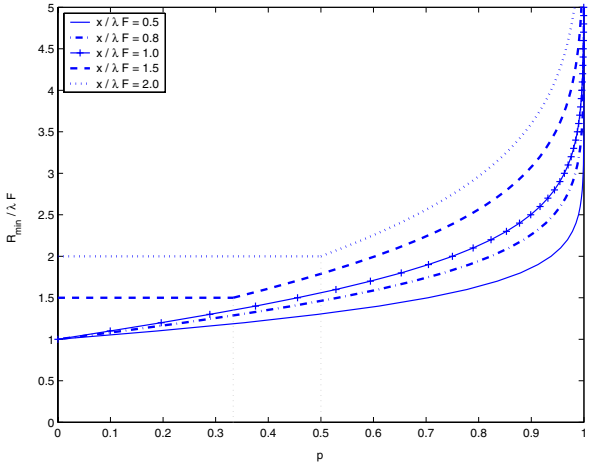

Fig. 3. R_{min} versus p

It follows that $G(x, R)$ increases monotonically with R for a fixed x. From (17), in order to satisfy $G(x, R) \geq p$, the rate $R \geq R_{min}$, where R_{min} is determined by the fixed point equation:

$$p = 1 - \left(\frac{\lambda F}{R_{min}}\right)^{R_{min}/x}, R_{min} \geq x, R_{min} > \lambda F \quad (19)$$

R_{min} versus p is shown in figure 3, where each curve represents a constant value of $x/\lambda F$. When $x \leq \lambda F$, R_{min} starts at λF when p = 0, increases monotonically with p, and approaches infinity as p approaches 1. When $x > \lambda F$, $R_{min} = x$ over $0 < p \leq 1 - \lambda F/x$, increases monotonically for higher values of p, and approaches infinity as p approaches 1.

We use the term *p-limited* to denote the case in which R_{min} increases monotonically with p, and the term *x-limited* to denote the case in which $R_{min} = x$. In the x-limited case, each user wants a guarantee of obtaining the full system transmission rate a certain portion of time. This can only occur when there is one user in the system, which

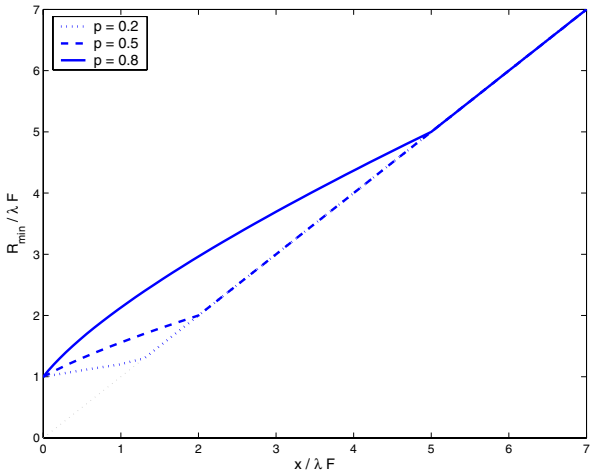

Fig. 4. R_{min} versus x

occurs with probability $1 - \lambda F/R$, and hence the x-limited case only occurs when $p \leq 1 - \lambda F/x$.

It follows that the marginal bandwidth required for an increase in the probability p of obtaining a rate of x or better is given by:

$$\frac{\partial R_{min}}{\partial p} = \left(\frac{R_{min}}{\lambda F}\right)^{R_{min}/x} \frac{x}{1 + \ln(R/\lambda F)}, \quad R_{min} \geq x, R_{min} > \lambda F$$

Finally, we examine the relationship between R_{min} and x. From (19), x can be represented as:

$$x = \min\left(\frac{R_{min}\ln(R_{min}/\lambda F)}{\ln(1/(1-p))}, R_{min}\right), R_{min} > \lambda F$$

It follows that the marginal bandwidth required for an increase in x at a fixed probability p is given by:

$$\frac{\partial R_{min}}{\partial x} = \begin{cases} \frac{\ln(1/(1-p))}{1+\ln(R_{min}/\lambda F)}, & \lambda F < R_{min} < \lambda F/(1-p) \\ 1, & R_{min} \geq \lambda F/(1-p) \end{cases}$$

R_{min} versus x is shown in figure 4, where each curve represents a constant value of p. R_{min} is monotonically increasing with x. When $x < \lambda F/(1-p)$, R_{min} is increasing with p. This is the p-limited case discussed above. When $x \geq \lambda F/(1-p)$, the system is x-limited, and $R_{min} = x$.

IV. Conclusion

In this paper, we introduced three definitions of average rate per job in a M/M/1 processor-sharing queue, m_{time}, $m_{weighted}$ and m_{jobs}. We proved that $m_{time} > m_{jobs} > m_{weighted}$ over all loads. Further we proved that the required system rate R_{min} is a monotonically increasing and convex function of the minimum average rate per jobs m. Finally, we showed that under a constraint on the tail probability of the average rate per job, the required system rate might be limited by either the probability p or by the location of the rate requirement x, and gave conditions explaining when each case occurs.

References

[1] E. G. Coffman, R. R. Muntz, and H. Trotter, "Waiting time distributions for processor-sharing systems," *J. ACM*, vol. 17, no. 1, pp. 123–130, Jan. 1970.

[2] J. A. Morrison, "Response-time distribution for a processor-sharing system," *SIAM J. Appl. Math.*, vol. 45, no. 1, pp. 152–167, Feb. 1985.

[3] F. Guillemin and J. Boyer, "Analysis of the M/M/1 queue with processor sharing via spectral theory," *Queueing System*, vol. 39, pp. 377–397, 2001.

[4] B. Sengupta and D. L. Jagerman, "A conditional response time of the M/M/1 processor-sharing queue," *ATT Tech. J.*, vol. 64, pp. 409–422, 1985.

[5] J. A. Morrison, "Aysmptotic analysis of the waiting-time distribution for a large closed processor-sharing system," *SIAM J. Appl. Math.*, vol. 46, pp. 140–170, 1986.

[6] ——, "Conditioned response-time distribution for a large closed processor-sharing system in very heavy usage," *SIAM J. Appl. Math.*, vol. 47, pp. 1117–1129, 1987.

[7] C. Knessl, "On the sojourn time distribution in a finite capacity processor shared queue," *J. ACM*, vol. 40, pp. 1238–1301, 1993.

[8] D. P. Gaver and P. A. Jacobs, "Processer-shared time sharing models in heavy-traffic," *SIAM J. Comput.*, vol. 15, pp. 1085–1100, 1986.

[9] A. Bersani and C. Sciarretta, "Asymptotic analysis for a closed processor-sharing system with switching times: Normal usage," *SIAM J, Appl. Math.*, vol. 51, pp. 525–541, 1991.

[10] R. Barbagallo, M. Mochi, and F. Zirilli, "Aysmptotic expansion of the waiting time distribution of two models of a closed processor-sharing system: Heavy usage," *SIAM J. Appl. Math.*, vol. 54, pp. 1468–1491, 1994.

[11] C. Knessl, "On the diffusion approximation to two parallel queues with processor sharing," *IEEE Trans. Automat. Control*, vol. 36, pp. 1356–1367, 1991.

[12] J. A. Morrison, "Diffusion approximation for head-of-the-line processor sharing for two parallel queues," *SIAM J. Appl. Math.*, vol. 53, pp. 471–490, 1993.

[13] V. Ramaswami, "The sojourn time in the GI/M/1 queue with processor sharing," *J. Appl. Prob.*, vol. 21, pp. 437–442, 1984.

[14] D. L. Jagerman and B. Sengupta, "The GI/M/1 processor-sharing queue and its heavy traffic analysis," *Stoch. Mod.*, vol. 7, pp. 379–395, 1991.

[15] B. Sengupta, "An approximation for the sojourn-time distribution for the GI/G/1 processor-sharing queue," *Stoch Mod.*, vol. 8, pp. 35–57, 1992.

[16] S. Grishechkin, "GI/G/1 processor sharing queue in heavy traffic," *Adv. Appl. Prob.*, vol. 26, pp. 539–555, 1994.

[17] Y. Yang and C. Knessl, "Conditional sojourn time moments in the finite capacity GI/M/1 queue with processor-sharing service," *SIAM J. Appl. Math.*, vol. 53, pp. 1132–1193, 1993.

[18] T. J. Ott, "The sojourn-time distribution in the M/G/1 queue with processor sharing," *J. Appl. Prob.*, vol. 21, pp. 360–378, 1984.

[19] S. F. Yashkov, "A derivation of response time distribution for a M/G/1 processor-sharing queue," *Problems of Control and Information Theory*, vol. 12, pp. 133–148, 1983.

[20] H. Masuyama and T. Takine, "Sojourn time distribution in a MAP/M/1 processor-sharing queue," *Operations Research Letters*, vol. 31, pp. 406–412, 2003.

[21] M. Harchol-Balter, K. Sigman, and A. Wierman, "Asymptotic convergence of scheduling policies with respect to slowdown," *Proc. of IFIP Perform.*, pp. 241–256, 2002.