REFERENCES

[1] T. J. Su and C. G. Huang, "Robust stability of delay dependence for linear uncertain systems," *IEEE Trans. Autom. Control*, vol. 37, no. 10, pp. 1656–1659, Oct. 1992.

[2] Y. Y. Cao, Y. X. Sun, and C. W. Cheng, "Delay-dependent robust stabilization of uncertain systems with multiple state delays," *IEEE Trans. Autom. Control*, vol. 43, no. 11, pp. 1608–1612, Nov. 1998.

[3] P. Park, "A delay-dependent stability criterion for systems with uncertain time-invariant delays," *IEEE Trans. Autom. Control*, vol. 44, no. 4, pp. 876–877, Apr. 1999.

[4] Y. S. Moon, P. Park, W. H. Kwon, and Y. S. Lee, "Delay-dependent robust stabilization of uncertain state-delayed systems," *Int. J. Control*, vol. 74, no. 14, pp. 1447–1455, 2001.

[5] E. Fridman and U. Shaked, "An improved stabilization method for linear time-delay systems," *IEEE Trans. Autom. Control*, vol. 47, no. 11, pp. 1931–1937, Nov. 2002.

[6] E. Fridman and U. Shaked, "Delay-dependent stability and H_∞ control: Constant and time-varying delays," *Int. J. Control*, vol. 76, no. 1, pp. 48–60, 2003.

[7] H. Gao and C. Wang, "Comments and further results on 'a descriptor system approach to H_∞ control of linear time- delay systems'," *IEEE Trans. Autom. Control*, vol. 48, no. 3, pp. 520–525, Mar. 2003.

[8] J.-P. Richard, "Time-delay systems: An overview of some recent advances and open problems," *Automatica*, vol. 39, no. 10, pp. 1667–1694, 2003.

[9] H. Gao, J. Lam, C. Wang, and Y. Wang, "Delay-dependent output-feedback stabilisation of discrete-time systems with time-varying state delay," *Proc. Inst. Elect. Eng Control Theory Appl.*, vol. 151, pp. 691–698, 2004.

[10] Q. L. Han, "On robust stability of neutral systems with time-varying discrete delay and norm-bounded uncertainty," *Automatica*, vol. 40, no. 6, pp. 1087–1092, 2004.

[11] Y. S. Lee, Y. S. Moon, W. H. Kwon, and P. G. Park, "Delay-dependent robust H_∞ control for uncertain systems with a state-delay," *Automatica*, vol. 40, no. 1, pp. 65–72, 2004.

[12] X. J. Jing, D. L. Tan, and Y. C. Wang, "An LMI approach to stability of systems with severe time-delay," *IEEE Trans. Autom. Control*, vol. 49, no. 7, pp. 1192–1195, Jul. 2004.

[13] Y. He, M. Wu, J. H. She, and G. P. Liu, "Parameter-dependent Lyapunov functional for stability of time-delay systems with polytopic-type uncertainties," *IEEE Trans. Autom. Control*, vol. 49, no. 5, pp. 828–832, May 2004.

[14] M. Wu, Y. He, J. H. She, and G. P. Liu, "Delay-dependent criteria for robust stability of time-varying delay systems," *Automatica*, vol. 40, no. 8, pp. 1435–1439, 2004.

[15] Y. He, M. Wu, J. H. She, and G. P. Liu, "Delay-dependent robust stability criteria for uncertain neutral systems with mixed delays," *Syst. Control Lett.*, vol. 51, no. 1, pp. 57–65, 2004.

[16] M. Wu, Y. He, and J. H. She, "New delay-dependent stability criteria and stabilizing method for neutral systems," *IEEE Trans. Autom. Control*, vol. 49, no. 12, pp. 2266–2271, Dec. 2004.

[17] S. Xu, J. Lam, and Y. Zou, "Simplified descriptor system approach to delay-dependent stability and performance analyses for time-delay systems," *Proc. Inst. Elect. Eng. Control Theory Appl.*, vol. 152, no. 2, pp. 147–151, 2005.

[18] C. Lin, Q.-G. Wang, and T. H. Lee, "A less conservative robust stability test for linear uncertain time-delay systems," *IEEE Trans. Autom. Control*, vol. 51, no. 1, pp. 87–91, Jan. 2006.

[19] X. M. Zhang, M. Wu, J. H. She, and Y. He, "Delay-dependent stabilization of linear systems with time-varying state and input delays," *Automatica*, vol. 41, no. 8, pp. 1405–1412, 2005.

[20] X. Jiang and Q. L. Han, "H_∞ control for linear systems with interval time-varying delay," *Automatica*, vol. 41, no. 12, pp. 2099–2106, 2005.

[21] Y. He, Q.-G. Wang, C. Lin, and M. Wu, "Augmented Lyapunov functional and delay-dependent stability criteria for neutral systems," *Int. J. Robust Nonlinear Control*, vol. 15, no. 18, pp. 923–933, 2005.

[22] J. K. Hale and S. M. Verduyn Lunel, *Introduction to Functional Differential Equations (Applied Mathematical Sciences*. New York: Springer-Verlag, 1993, vol. 99.

[23] P. Pepe, The problem of the absolute continuity for Lyapunov–Krasovskii Functionals University of L'Aquila, L'Aquila, Italy, Res. Rep. R.05-80 [Online]. Available: http://www.diel.univaq.it/research/

Throughput in Processor-Sharing Queues

Na Chen and Scott Jordan

***Abstract*—Processor-sharing queues are often used to model file transmission in networks. While sojourn time is a common performance metric in the queueing literature,** *average transmission rate* **is the more commonly discussed metric in the networking literature. Whereas much is known about sojourn times, there is little known about the average service rate experienced by jobs in processor-sharing queues. We first define the average rate as observed by users and by the queue. In an M/M/1 processor-sharing queue, we give closed-form expressions for these average rates, and prove a strict ordering amongst them. We prove that the queue service rate (in bps) is an increasing function of the minimum required average transmission rate, and give a closed-form expression for the marginal cost associated with such a performance requirement. We then consider the effect of using connection access control by modeling an M/M/1/K processor-sharing queue. We give closed-form expressions for average transmission rates, and discuss the relationship between the queue service rate (in bps), the queue limit, the average rate, and the blocking probability.**

***Index Terms*—Average rate, marginal cost, processor-sharing (PS) queues.**

I. INTRODUCTION

We are motivated here by the idea of providing performance guarantees for elastic data applications. In particular, we consider minimum bounds on the mean transmission rate or throughput. Intuition suggests that the cost of providing such a guarantee should be increasing with the level of the guarantee, but this intuition has not yet been grounded with a theoretical basis. In this paper, we consider such guarantees in the context of processor-sharing queues.

Processor-sharing (PS) queues have been widely used in the networking literature to model multiple file transmissions dynamically sharing a fixed amount of bandwidth [1]–[3]. Each user or job represents transmission of a file, and the queue service rate (in bps) represents the bandwidth of the system. A user starts transmission when it arrives and departs when the file transmission has completed. The number of active users is stochastic, and so is the transmission rate per user under the PS discipline.

The processor-sharing service discipline is an appropriate model when the time scale of interest is call-level and all files share bandwidth equally [4], [5]. The call-level time scale applies when the relevant performance metrics are measured over the typical length of a file transmission; if the relevant metrics are measured on a packet-level time scale, then the scheduler is usually modeled as one that swaps between jobs. The equal bandwidth assumption is often made when there exists a mechanism, e.g., the transmission control protocol (TCP), that attempts to equalize bandwidth between multiple streams over multiple round trip times.

There is rich literature concerning processor-sharing queues. The most common call-level performance metric for such queues is sojourn time. When modeling file transmission, sojourn time corresponds to the time required to completely transmit the file, which is certainly

Manuscript received November 15, 2005; revised August 9, 20006. Recommended by Associate Editor I. Paschalidis. This work was supported by the National Science Foundation under Grant CNS-0137103 and by the Defense Advanced Research Projects Agency under Grant N66001-00-8935.

The authors are with the Department of Electrical Engineering and Computer Science, the University of California, Irvine, CA 92697 USA (e-mail: nac@uci.edu; sjordan@uci.edu).

Digital Object Identifier 10.1109/TAC.2006.887906

of interest. For M/M/1-PS queues, Coffman [6] 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. By removing the conditioning and inverting this Laplace transform, Morrison [7] obtained an integral representation for the complementary distribution of the sojourn time, which was refined by Guillemin [8] 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. For M/M/1/K-PS queues, Morrison obtained an asymptotic approximation to the equilibrium distribution of the waiting time [9]. In heavy-traffic, Morrison [10] also found the distribution of the response time conditioned on the service time, and Knessl [11] constructed an asymptotic approximation to the sojourn time distribution.

However, the most common performance metric for data applications in the Internet is throughput, not sojourn time. Indeed, many Internet service providers advertise a speed of some type when selling residential broadband service, and there are many online speed tests that measure throughput on a broadband connection. Throughput is casually perceived as the rate at which a computer or network sends or receives data. However, a more precise definition of throughput is with respect to a time window, as the number of bits transmitted divided by the length of the time window. The time window is traditionally chosen to correspond to the time scale on which users judge performance, typically ranging from tenths of a second for highly interactive applications such as gaming, to seconds for moderately interactive applications such as web browsing, to minutes for noninteractive applications such as file downloads.

While there are many available results on sojourn time in process-sharing queues, there is little literature on the throughput in such queues. Definitions of average rate as observed by users (here called the *average rate over jobs*) and of average rate as observed by the queue (here called the *average rate over time*) were introduced in [5]. Other common metrics include slowdown (cf. [12]), mean slowdown (cf. [13]), and flow throughput (cf. [1] and [2]). A closed-form expression for the average rate over time has been derived, and in a G/D/1-PS queue it has been proven that the average rate over time dominates the average rate over jobs [5].

In Section II, we present the definitions of average rates, which has also been introduced by [5]. Then we start in Section III by deriving a closed-form expression for the average rate over jobs, and by proving that in an M/M/1-PS queue the average rate over time dominates the average rate over jobs. In Section IV, we then consider the cost associated with providing performance guarantees on the mean rate, by examining the marginal queue service rate with respect to the minimum required average rate over time. We prove that the cost is monotonically increasing with both the average rate over jobs and the average rate over time, consistent with intuition. In addition, we prove convexity of the marginal costs associated with average rate over time, and derive asymptotic behavior of costs for the average rates over both jobs and time.

In Section V, we then turn to the effect of connection access control upon average rates, which we do not believe has been addressed in the literature. By considering an M/M/1/K-PS queue, we present similar definitions for average rates over both jobs and time. The relationships between queue service rate, the queue limit, the average rate over time, and the blocking probability are investigated, and we demonstrate the nature of binding constraints on mean rate and on blocking probability.

II. Performance Measures

In this section, we present definitions of average rate as observed by users and by the queue in a G/G/1 processor-sharing queue. It is assumed that the queue is ergodic.

From the queue's perspective, the bandwidth is split among all jobs present in the system and, therefore, the instantaneous rate per job changes whenever a job arrives or departs. Let R (bps) denote the bandwidth, $n(u)$ the number of jobs in the queue at time u. Then, the instantaneous transmission rate received by each job at time u is given by $(R/n(u))I_{\{n(u)>0\}}$, where $I_{\{\cdot\}}$ is the indicator function.

An average rate as observed by the queue can be defined by averaging the instantaneous rate over time conditioned on at least one job in the system. We call this quantity the *average rate over time*, m^T, given by

$$m^T \equiv \lim_{t\to\infty} \frac{\int_0^t \frac{R}{n(u)} I_{\{n(u)>0\}}\,du}{\int_0^t I_{\{n(u)>0\}}\,du} = E\left[\frac{R}{N}\middle| N>0\right] \tag{1}$$

$E[\cdot]$ denotes the expectation of the quantity within square brackets. N is a random variable, denoting the number of jobs in the system.

A second definition of average rate as observed by the queue could be created by weighting the instantaneous rate by the number of jobs receiving that rate. We define the *weighted average rate over time*, m^W, as

$$m^W \equiv \lim_{t\to\infty} \frac{\int_0^t n(u)\frac{R}{n(u)} I_{\{n(u)>0\}}\,du}{\int_0^t n(u) I_{\{n(u)>0\}}\,du} = \frac{R(1-\pi_0)}{E[N]} \tag{2}$$

where π_0 denotes the probability that the queue is empty. If the job size and the sojourn time are denoted by L and T, respectively, then it can be shown that using Little's law m^W is equal to $E[L]/E[T]$, which has also been called the *flow throughput* [2].

From the user's perspective, the average transmission rate (or throughput) r_i for job i is defined as its file size divided by the time required to transmit the file. An average rate as observed by users can be defined by averaging this quantity over the users. We define the *average rate over jobs*, m^J, as the expected throughput per job, given by

$$m^J \equiv \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} r_i = E_i[r_i] = E\left[\frac{L}{T}\right]. \tag{3}$$

The average rate over time m^T and weighted average rate over time m^W rely on the stationary distribution of the queue, which is usually easy to compute. In contrast, the average rate over jobs m^J depends on the distribution of sojourn time conditional on job size, which is complicated in most cases. The relationship among these average rates is also of interest.

Theorem 1: In a G/G/1-PS queue, $m^T \geq m^W$.

Proof: Divide m^T in (1) by m^W in (2) and apply Jensen's inequality:

$$\frac{m^T}{m^W} = \frac{E[\frac{1}{N}|N>0]E[N]}{1-\pi_0} \geq \frac{1}{1-\pi_0} \geq 1$$

The theorem follows. ∎

Our intuition is that $m^T \geq m^W$ follows from the convexity of $1/N$, which results in $E[1/N] > 1/E[N]$. It has also been shown that $m^J \geq m^W$ in an M/G/1-PS queue and $m^T > m^J$ in a G/D/1-PS queue [5]. For other distributions of job size, the relationship between m^T and m^J is still unknown. We will next address this issue in an M/M/1-PS queue.

III. Average Rates in an M/M/1-PS Queue

In this section, we give closed-form expression for each of these definitions for average rate in an M/M/1-PS queue, and prove that the average rate over jobs m^J dominates the average rate over time m^T.

Consider an M/M/1-PS queue in which jobs arrive as a Poisson process with rate λ (jobs/s) and file sizes are i.i.d. Exponentially distributed random variables with mean l (bits). The service rate of the queue (in jobs/s) is denoted by $\mu = R/l$, and then the offered load is denoted by $\rho \equiv \lambda/\mu$. We assume that the queue is ergodic, namely $\rho < 1$. It follows that the stationary distribution of the number of jobs in the queue is given by $\pi_n \equiv \Pr(N = n) = (1-\rho)\rho^n, n = 0, 1, \ldots$[1] [14].

We now start to derive closed-form expressions for the average rates defined previously.

Theorem 2: In an M/M/1-PS queue, the average rate over time is given by

$$m^T = \frac{R(1-\rho)}{\rho} \ln \frac{1}{1-\rho} \tag{4}$$

the weighted average rate over time is given by

$$m^W = R(1-\rho). \tag{5}$$

Proof: m^T in (1) is equal to $\sum_{n=1}^{\infty}(R\pi_n/(n(1-\pi_0)))$. Substituting π_n and expressing the sum as logarithm yields (4). It is readily shown that $E[N] = \rho/(1-\rho)$ and, thus, substituting $E[N]$ and π_0 in (2) yields (5). ∎

Due to the *insensitivity property* of PS queues, these two expressions also hold for an M/G/1-PS queue.

The derivation of a closed-form expression for the average rate over jobs is more involved, since it depends on the ratio of a user's service requirement to the user's sojourn time, and therefore relies on the distribution of sojourn time conditional on job size.

Theorem 3: In an M/M/1-PS queue, the average rate over time is given by:

$$m^J = R(1-\rho)\int_0^{\infty} v e^{-v}\left(\int_0^{\infty} c_1(\rho,u,v)\,du\right)dv \tag{6}$$

where

$$c_1(\rho,u,v) = \frac{(1-\rho r^2)e^{-\frac{1-r}{r}v}}{(1-\rho r)^2 - \rho(1-r)^2 e^{-\frac{(1-\rho r^2)}{r}v}},$$

and r is the smaller root of the equation $\rho r^2 - (1+u+\rho)r + 1 = 0$.

Proof: Denote the file size and sojourn time of job i by L_i and T_i correspondingly, so that its throughput is given by $r_i = L_i/T_i$. Due to the arrivals from 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, 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 user's job size, or equivalently upon its service time requirement, which is equal to the job size divided by the total rate. Denote the throughput of the job with size $R\tau$ (and, hence, service time requirement τ) by

$$Y(\tau) = E_i[r_i | L_i = R\tau] = R\tau E_i\left[\frac{1}{T_i}\middle| L_i = R\tau\right].$$

It follows from (3) that $m^J = E_\tau[Y(\tau)]$.

An expression for $E_i[1/T_i | L_i = R\tau]$ can be found by integrating the conditional Laplace transform of sojourn time conditioned on a job's service time, $E_i[e^{-sT_i}|L_i = R\tau]$, given in [6]. Thus, the average rate for jobs with service time τ is given by

$$\begin{aligned} Y(\tau) &= R\tau\int_0^{\infty} E_i[e^{-sT_i}|L_i = R\tau]\,ds \\ &= R\tau\int_0^{\infty} c_2(\lambda,\mu,s,\tau)\,ds \end{aligned}$$

where

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

and r is the smaller root of the equation $\lambda r^2 - (\lambda+\mu+s)r + \mu = 0$.

Furthermore, since L_i is Exponentially distributed with mean l, the average rate as seen by users is

$$\begin{aligned} m^J &= \int_0^{\infty} \frac{R\tau}{l} e^{-\frac{R\tau}{l}}\left(\int_0^{\infty} c_2(\lambda,\mu,s,\tau)\,ds\right)d(R\tau) \\ &= \frac{R^2}{l}\int_0^{\infty} \tau e^{-\frac{R\tau}{l}}\left(\int_0^{\infty} c_2(\lambda,\mu,s,\tau)\,ds\right)d\tau. \end{aligned}$$

An equivalent expression has been independently found by [15]; however, we wish to obtain an expression only in terms of the bandwidth R and the offered load ρ.

In order to write m^J solely in terms of R and ρ, we use $s = (1-r)(1-\rho r)\mu/r$ given in [6]. Define $u = s/\mu = (1-r)(1-\rho r)/r$ and $v = \mu\tau$. Substituting R/F by μ and using a variable substitution from $\{s,\tau\}$ to $\{u,v\}$, we obtain (6). ∎

We now turn to a comparison of m^T and m^J, as discussed before. Our principal result is given in the following theorem.

Theorem 4: In an M/M/1-PS queue, $m^T > m^J$, for all $0 < \rho < 1$.

Proof: We rewrite m^T as represented in (7) and compare it with m^J represented in (6). We start by expressing the term $\ln(1/(1-\rho))$ as a double integral with respect to u and v

$$\begin{aligned} \ln\frac{1}{1-\rho} &= \ln\frac{1}{1-\rho}\int_0^{\infty} e^{-v}\,dv \\ &= \int_0^{\infty} e^{-v}[\ln(1-\rho e^{-uv})]_{u=0}^{u=\infty}\,dv \\ &= \int_0^{\infty} \rho v e^{-v}\left(\int_0^{\infty}\frac{e^{-uv}}{1-\rho e^{-uv}}\,du\right)dv. \end{aligned}$$

Then m^T in (4) can be written in a similar form as m^J in (6):

$$m^T = R(1-\rho)\int_0^{\infty} v e^{-v}\left(\int_0^{\infty} c_4(\rho,u,v)\,du\right)dv \tag{7}$$

where $c_4(\rho,u,v) = e^{-uv}/(1-\rho e^{-uv})$. Comparing (7) with (6), it follows that a sufficient condition for $m^T > m^J$ to hold is $c_4(\rho,u,v) > c_1(\rho,u,v)$, for all $\{u,v\} > 0$ and $0 < \rho < 1$.

[1]Processor-sharing queues have the well-known *insensitivity property*, i.e., the queue stationary distribution is independent of the distribution of the service requirement. Therefore, an M/G/1-PS queue has the same stationary distribution.

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

$$(1-\rho r)^2 e^{\frac{(1-\rho r^2)}{r}v} > \rho(1-r)^2 + (1-\rho r^2)\left[e^{\frac{(1-\rho r)}{r}v} - \rho e^{(1-\rho r)v}\right].$$

Finally, using the Maclaurin series of e^x and simplifying, it follows that $m^T > m^J$ if

$$\sum_{n=3}^{\infty} \frac{v^n(1-\rho r^2)(1-\rho r)^2}{n! r^n} H_n(\rho, r) > 0$$

where $H_n(\rho, r) = (1-\rho r^2)^{n-1} - (1-\rho r)^{n-2}(1-\rho r^n)$.

We will show that the sum is positive by showing that each term is positive, using mathematical induction. The base case is $H_3(\rho, r) = \rho r(1-r)^2$. Since $0 < \rho < 1$ and $0 < r < 1$ (which is fairly simple to prove), it follows that $H_3(\rho, r) > 0$. For the induction case, assume that $H_k(\rho, r) > 0$ for some $k > 3$, that is, $(1-\rho r^2)^{k-1} > (1-\rho r)^{k-2}(1-\rho r^k)$

$$\begin{aligned} &H_{k+1}(\rho, r) \\ &\quad > (1-\rho r)^{k-2}(1-\rho r^k)(1-\rho r^2) \\ &\qquad - (1-\rho r)^{k-1}(1-\rho r^{k+1}) \\ &\quad = \rho r(1-\rho r)^{k-2}(1-r)(1-r^{k-1}). \end{aligned}$$

Since $r < 1$, it follows that $H_{k+1}(\rho, r) > 0$. Consequently, $H_n(\rho, r) > 0$, for $n \geq 3$, and hence $m^T > m^J$. ■

Our intuition is that $m^J < m^T$ is due to the effect of long jobs. To understand this, consider the two extremes in terms of job length—a job with an infinitesimal file size and a job with an infinite file size. Due to Poisson arrivals, a new job will see upon arrival, a distribution of jobs in the system (excluding itself) that is the same as the stationary distribution of the system. A job with an infinitesimal file size will experience no fluctuation in rate during its sojourn time; a simple calculation shows that this job on average experiences a rate equal to m^T. A job with an infinite file size, however, would experience over its sojourn time a different stationary distribution; a simple calculation shows that this job experiences an average rate equal to m^W, which is strictly lower than m^T when $0 < \rho < 1$. Jobs with intermediate lengths also experience an average rate strictly lower than m^T when $0 < \rho < 1$. As a consequence, m^J, which averages jobs of different lengths, is strictly lower than m^T when $0 < \rho < 1$.

The three average rates (normalized by R) are plotted versus the load ρ in Fig. 1. We observe that the difference between m^T and m^J is relatively small, and that the difference between m^T or m^J and m^W is concave with respect to ρ.

IV. Performance Guarantees in an M/M/1-PS Queue

In this section, we consider performance guarantees on the average rate over time and on the average rate over jobs, expressed as $m^T \geq m$ and $m^J \geq m$ respectively with m denoting the minimum required average rate. We prove that the cost of bandwidth is monotonically increasing with m in both cases, consistent with intuition. In addition, we prove convexity of the marginal costs associated with average rate over time, and derive asymptotic behavior of costs for the average rates over both time and jobs.

We start by considering the performance guarantee of the form $m^T \geq m$. For purposes of discussion, we assume that the user arrival rate λ and the mean file size l are fixed, but that the bandwidth R is chosen to satisfy the performance bound. We expect that satisfaction of $m^T \geq m$ thus requires that R exceed a related lower bound, denoted by $R^T_{\min}$.

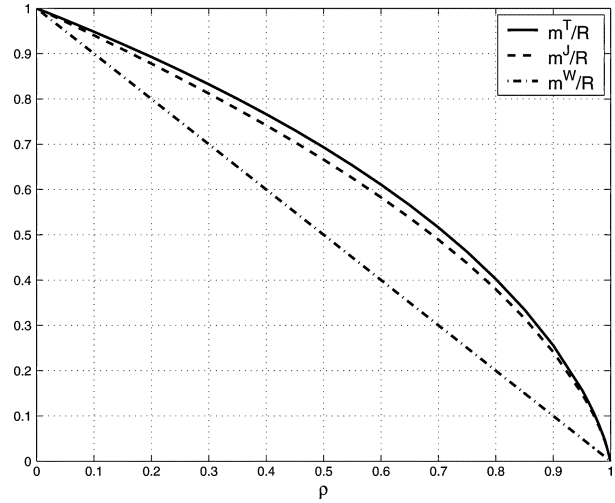

Fig. 1. Average rates in an M/M/1-PS queue.

Our first result is a closed-form expression for the marginal bandwidth $R^T_{\min}$ with respect to the minimum required average rate m.

Theorem 5: In an M/M/1-PS queue, with $\rho_{\max} = \lambda l / R^T_{\min}$

$$\frac{\partial R^T_{\min}}{\partial m} = \left(\frac{2-\rho_{\max}}{\rho_{\max}} \ln \frac{1}{1-\rho_{\max}} - 1\right)^{-1}. \tag{8}$$

Proof: The derivative of m^T with respect to R, following from (4):

$$\frac{\partial m^T}{\partial R} = \frac{2-\rho}{\rho} \ln \frac{1}{1-\rho} - 1 > 1 - \rho > 0. \tag{9}$$

The first inequality comes from $\ln(1/(1-\rho)) > \rho$. Hence m^T increases monotonically with R. To satisfy the performance guarantee $m^T \geq m$, the bandwidth $R \geq R^T_{\min}$, where $R^T_{\min}$ is determined by the fixed point equation $m^T(R) = m$. Thus, $\partial R^T_{\min}/\partial m = 1/(\partial m^T/\partial R)|_{\{R=R^T_{\min}, m^T=m\}}$. The theorem follows. ■

This relationship can be used to guide dimensioning algorithms, as it relates the minimum required bandwidth $R^T_{\min}$ to the arrival rate (in bits) λl and the minimum required average rate m. Furthermore, if a pricing policy is used that bases the price on the marginal cost, then the price would be determined by (8).

A stronger characterization of $R^T_{\min}$ versus m is described in the following theorem.

Theorem 6: $R^T_{\min}$ is a monotonically increasing and convex function of m, and $R^T_{\min} - m$ monotonically decreases with m from λl to $\lambda l/2$.

Proof: From (9), $(\partial m/\partial R^T_{\min}) > 0$. Thus, $R^T_{\min}$ monotonically increases from λl to ∞ as m increases from 0 to ∞. Consider the second derivative

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

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

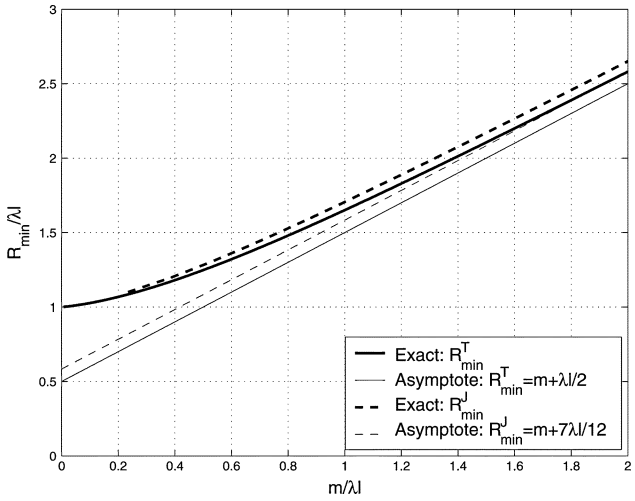

Fig. 2. $R^T_{\min}$ and $R^J_{\min}$ versus m in an M/M/1-PS queue.

$f'(\rho_{\max}) = 2\rho_{\max} + 2\ln(1-\rho_{\max}) < 0$. Hence, the second derivative is negative and thus $R^T_{\min}$ is a convex function of m.

To prove the variation of $R^T_{\min} - m$ with m, expand $\ln(1-\rho_{\max})$ by its Maclaurin series, substituting in $m^T(R^T_{\min}) = m$, and simplifying yields

$$R^T_{\min} - m = \left(\frac{1}{2} + \sum_{n=1}^{\infty} \frac{\rho^n_{\max}}{(n+1)(n+2)}\right)\lambda l.$$

When $\rho_{\max}$ decreases monotonically from 1 to 0, $R^T_{\min} - m$ decreases monotonically from λl to $\lambda l/2$. As m increases from 0 to ∞, we have already noted that $R^T_{\min}$ increases monotonically from λl to ∞. It follows that $R^T_{\min} - m$ decreases monotonically from λl to $\lambda l/2$ as m increases. ■

If the cost of providing a performance guarantee of the form $m^T \geq m$ is proportional to $R^T_{\min}$, then it follows from this theorem that the price would be increasing and convex in m.

$R^T_{\min}$ versus m is shown in Fig. 2, along with its asymptote $R^T_{\min} = m + \lambda l/2$. The asymptote can be thought of as a limit of the required rate as the load approaches zero. In the limit, during a busy cycle there is one job in the system with probability $1-\rho$ and two jobs with probability ρ, neglecting $O(\rho^2)$ terms. It follows that the corresponding limit is given by $m^T = (1-\rho)R + \rho R/2 = R - \lambda l/2$.

We now consider the performance guarantee on the average rate over jobs, i.e., of the form $m^J \geq m$. We would expect that satisfaction of $m^J \geq m$ thus also requires that the bandwidth R exceed a related lower bound, denoted by $R^J_{\min}$. A similar characterization is given by the following theorem.

Theorem 7: $R^J_{\min}$ is a monotonically increasing function of m, and $R^J_{\min} - m$ monotonically decreases with m from λl to $7\lambda l/12$.

Proof: Under a fixed system load, an increase in the bandwidth R would lead to a corresponding decrease in all jobs' sojourn times, a corresponding increase in all jobs' throughputs, and thus a corresponding increase in the average rate over jobs m^J. It trivially follows that m^J monotonically increases with the bandwidth R, and thus that $R^J_{\min}$ increases monotonically with m.

It remains to prove the asymptote. As m approaches $0, R^J_{\min}$ approaches λl (required for ergodicity). As m approaches ∞, i.e., as $R^J_{\min}$ approaches ∞ and the load ρ approaches 0, from (6) the corresponding limit of $m/R^J_{\min}$ is given by

$$\lim_{\rho\to 0}\frac{m}{R^J_{\min}} = \int_0^\infty ve^{-v}\left(\int_0^\infty \left(\lim_{\rho\to 0} c_1(\rho,u,v)\right)du\right)dv = \int_0^\infty ve^{-v}\left(\int_0^\infty e^{-uv}\,du\right)dv = 1.$$

It follows that, as ρ approaches $0, R^J_{\min}$ linearly increases with m at slope 1. We now calculate the limit of $R^J_{\min} - m$ as ρ approaches 0

$$\begin{aligned}&\lim_{\rho\to 0}\left(R^J_{\min} - m\right)\\ &= \lambda l \lim_{\rho\to 0}\frac{1}{\rho}(1-(1-\rho)\\ &\quad\times\int_0^\infty ve^{-v}\left(\int_0^\infty c_1(\rho,u,v)\,du\right)dv\Big)\\ &= \lambda l\left(1 - \int_0^\infty ve^{-v}\right.\\ &\quad\left.\times\left(\int_0^\infty\left(\lim_{\rho\to 0}\frac{\partial c_1(\rho,u,v)}{\partial\rho}\right)du\right)dv\right).\end{aligned} \tag{10}$$

It can be shown that

$$\lim_{\rho\to 0}\frac{\partial c_1(\rho,u,v)}{\partial\rho} = -\left(\frac{1}{(u+1)^2} + \frac{uv-2}{u+1} - \frac{u^2e^{-(u+1)v}}{(u+1)^2}\right)e^{-uv}. \tag{11}$$

Substituting (11) in (10) and integrating yields $\lim_{\rho\to 0}(R^J_{\min} - m) = (7/12)\lambda l$. The lower asymptote in the theorem follows. ■

$R^J_{\min}$ versus m is also shown in Fig. 2, along with its asymptote $R^J_{\min} = m + 7\lambda l/12$. Observe that the bandwidth required by the guarantee on the average rate over jobs is higher than that required by same level guarantee on the average rate over time, consistent with their order stated in Theorem 4.

V. Average Rates and Performance Guarantees in an M/M/1/K-PS Queue

In an M/M/1-PS queue, the system provides a higher level of performance guarantees by increasing the bandwidth. However, in many practical situations, the available bandwidth is physically limited and can not be increased, or increasing bandwidth may not be the most efficient manner to satisfy the performance requirement. In this case, connection access control (CAC) is commonly used to maintain acceptable performance for admitted jobs, at the cost of blocking some jobs. This approach can be modeled by an M/M/1/K-PS queue, where K is the queue limit, i.e., the maximum number of jobs that can transmit simultaneously. CAC gives the network designer additional flexibility, by allowing for a tradeoff between the bandwidth and the queue limit.

In this section, we first present closed-form expressions of average rates and blocking probability in an M/M/1/K-PS queue. We then consider the ability of the system to provide a guarantee on average rate over time by appropriately choosing the bandwidth R or the queue limit K. We finally demonstrate the nature of binding constraints on the average rate over time and on the blocking probability.

The stationary distribution for the number of users in an M/M/1/K-PS queue is given by $\pi_n = \Pr(N = n) = (1-\rho)\rho^n/(1-\rho^{K+1}), n = 0, 1, \ldots, K$ [14].

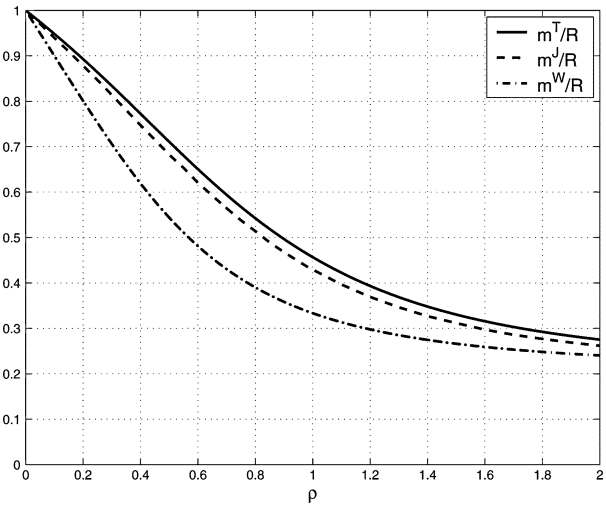

Fig. 3. Average rates in an M/M/1/5-PS queue.

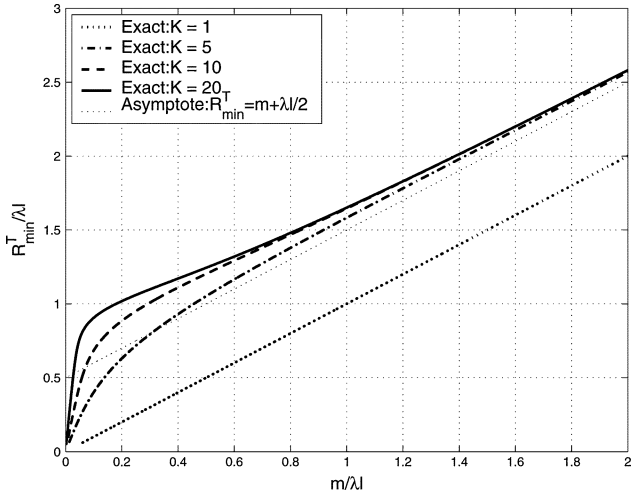

Fig. 4. $R^T_{\min}$ versus m in an M/M/1/K-PS queue.

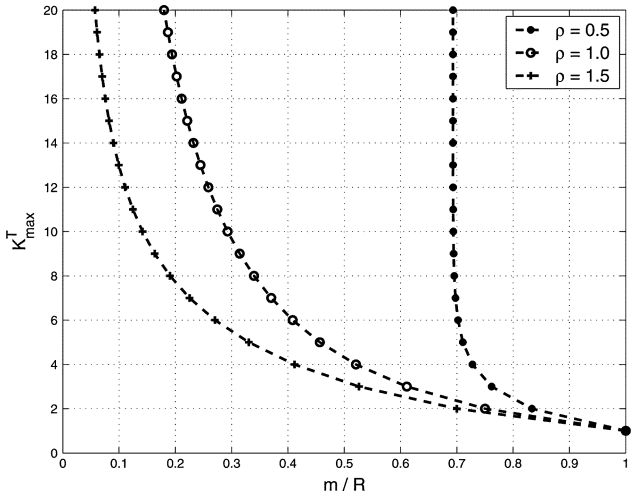

Fig. 5. $K^T_{\max}$ versus m in an M/M/1/K-PS queue.

Theorem 8: In an M/M/1/K-PS queue, the average rate over time is given by

$$m^T = \frac{R(1-\rho)}{1-\rho^K}\sum_{n=1}^{K}\frac{\rho^{n-1}}{n} \tag{12}$$

the weighted average rate over time is given by

$$m^W = \frac{R(1-\rho)(1-\rho^K)}{1-(K+1)\rho^K+K\rho^{K+1}}. \tag{13}$$

Proof: m^T in (1) is equal to $\sum_{n=1}^{K}(R\pi_n/(n(1-\pi_0)))$. Substituting π_n yields (12). It is readily shown that $E[N] = \rho(1-(K+1)\rho^K+K\rho^{K+1})/((1-\rho^{K+1})(1-\rho))$, and thus substituting $E[N]$ and π_0 in (2) yields (13). ■

Unfortunately, it appears difficult to obtain a closed-form expression for m^J. Fig. 3 shows average rates (normalized by R) versus ρ in the case $K = 5$, where m^T and m^W are calculated from the above formulas while m^J is obtained by simulation. Compared to Fig. 1, the order among them is same, that is, $m^T > m^J > m^W$ for $\rho > 0$. (We conjecture, but are unable to prove without the expression of m^J, that this order holds for all K.) Moreover, the difference between m^T and m^J is also relatively small compared to the difference between either of them and m^W. As the load approaches ∞, all normalized average rates approach a common lower bound equal to the minimum instantaneous rate, $1/K$.

The blocking probability is given by $P^B = \pi_K = (1-\rho)\rho^K/(1-\rho^{K+1})$. It is readily shown that P^B decreases as R increases or as K increases.

We start to consider the performance bound on the average rate over time, which is in the form $m^T \geq m$.

One approach to provide this guarantee is to increase the bandwidth R with K being fixed since the average rate over time is an increasing function of R. Fig. 4 shows the minimum required bandwidth $R^T_{\min}$ (normalized by λl) versus m for different values of K. Note that an M/M/1/K-PS queue is always ergodic, and hence $R^T_{\min}$ can be less than λl. When m approaches 0, i.e., no guarantee is needed, $R^T_{\min}$ also approaches 0; as m increases, $R^T_{\min}$ increases monotonically, as it did in the M/M/1-PS queue. For all $K \geq 2$, $R^T_{\min}$ asymptotically approaches the line $R^T_{\min} = m + \lambda l/2$ which is the same as the asymptote in the M/M/1-PS queue. ($K = 1$ is a special case, in which $R^T_{\min} = m$.) This occurs because as the load approaches 0, the stationary distribution of the M/M/1/K-PS queue converges to that of the M/M/1-PS queue.

An alternative manner to provide the guarantee $m^T \geq m$ is to leave the bandwidth R fixed and to decrease K until the guarantee is satisfied. This approach can be justified using the following theorem.

Theorem 9: m^T monotonically decreases with K for a fixed R.

Proof: Subtract $m^T(K+1)$ from $m^T(K)$

$$\begin{aligned} &m^T(K) - m^T(K+1) \\ &= \frac{R(1-\rho)^2\rho^K}{(1-\rho^K)(1-\rho^{K+1})} \\ &\quad\times\left(\sum_{n=1}^{K}\frac{\rho^{n-1}}{n} - \frac{1-\rho^K}{(K+1)(1-\rho)}\right) \\ &= \frac{R(1-\rho)^2\rho^K}{(1-\rho^K)(1-\rho^{K+1})} \\ &\quad\times\left(\sum_{n=0}^{K-1}\frac{\rho^n}{n+1} - \sum_{n=0}^{K-1}\frac{\rho^n}{K+1}\right) > 0. \end{aligned}$$

The theorem follows. ■

It follows that there exists an upper bound $K^T_{\max}$ such that $K \leq K^T_{\max}$ will result in satisfaction of $m^T \geq m$. Fig. 5 shows $K^T_{\max}$ versus

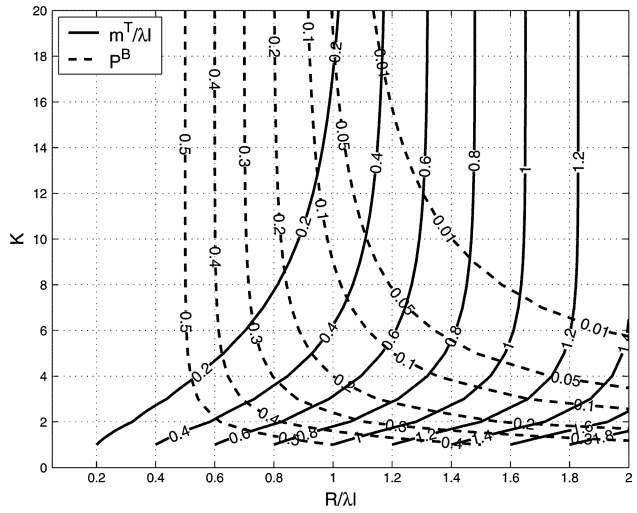

Fig. 6. K versus R for fixed m^T or P^B in an M/M/1/K-PS queue.

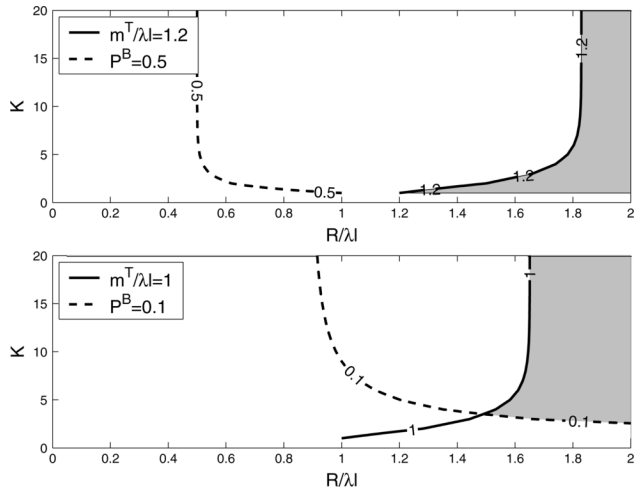

Fig. 7. Two cases of feasible regions for R and K.

m for different values of the system load ρ. When $m = R, K = 1$ for all load ρ.

We finally consider the joint selection of the bandwidth R and the queue limit K to jointly satisfy performance requirements on the average rate over time $m^T \geq m$ and on the blocking probability $P^B \leq p_b$. We are still interested in the relationship between the minimum required bandwidth $R^T_{\min}$ and the minimum required average rate m as well as the maximum required blocking probability p_b. Our main result here is given as follows: if $m/(\lambda l) \geq (1 - p_b)/p_b$, then $R^T_{\min} = m$ while $K = 1$; otherwise, $R^T_{\min}$ and K can be obtained by solving the equation set: $m^T = m$ and $P^B = p_b$.

Fig. 6 shows the contour lines of m^T (solid curves) and of P^B (dashed curves) on the surface specified by R (normalized by λl) and K; the attained value of m^T or P^B is denoted on each curve. As discussed previously, m^T increases with R and decreases with K, and thus K is increasing with R on each contour line of m^T; similarly, P^B decreases with both R and K and, thus, K is decreasing with R on each contour line of P^B.

There exists two possible cases of intersection for a solid curve and a dashed curve, which correspond to two cases for the feasible region of (R, K) given a pair of performance requirements (m, p_b). Fig. 7 illustrates these two types of feasible regions as shaded areas. In the first case, the set of $\{(R, K)|m^T \geq m\}$ is a subset of $\{(R, K)|P^B \leq p_b\}$. It is easily shown that this case occurs when $m/(\lambda l) \geq (1 - p_b)/p_b$, and the minimum bandwidth $R^T_{\min} = m$ with $K = 1$. The second case is that the set of $\{(R, K)|m^T \geq m\}$ partially overlaps with the set of $\{(R, K)|P^B \leq p_b\}$. The minimum bandwidth lies at the intersection of $m^T = m$ and $P^B = p_b$, namely where both constraints are binding.

Note that it is impossible that the set of $\{(R, K)|m^T \geq m\}$ includes the set of $\{(R, K)|P^B \leq p_b\}$. As K approaches ∞, the minimum bandwidth required by $P^B \leq p_b$ approaches 0, whereas the minimum bandwidth required by $m^T \geq m$ is always greater than 0.

VI. Conclusion

We have focused on the average transmission rate as a performance metric in processor-sharing queues. Whereas much is known about sojourn times in PS queues, little is known about average rates. We introduced three definitions of average rate as observed by users and by the queue. In an M/M/1-PS queue, we proved that the average rate over time dominates the average rate over jobs. By giving expressions for the marginal bandwidth with respect to average rates, we characterized the difficulty of guaranteeing minimum average rates. We considered the effect of connection access control, and characterized when performance bounds on average rate and/or blocking probability are binding.

We believe such results are useful in dimensioning processor-sharing queues when performance is measured by average rate or throughput. In particular, we expect that such results can be used within networking to design scheduling and connection access control policies for data services.

References

[1] J. W. Roberts, "A survey on statistical bandwidth sharing," *Comput. Networks*, vol. 45, no. 3, pp. 319–332, 2004.

[2] T. Bonald and L. Massouli, "Impact of fairness on Internet performance," *Proc. SIGMETRICS'01*, pp. 82–91, 2001.

[3] A. W. Berger and Y. Kogan, "Dimensioning bandwidth for elastic traffic in high-speed data networks," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 643–654, 2000.

[4] P. S. Chanda, A. Kumar, and A. A. Kherani, "An approximate calculation of max-min fair throughputs for non-persistent elastic flows," *Proc. GLOBECOM'01*, vol. 3, pp. 1892–1897, 2001.

[5] A. A. Kherani and A. Kumar, "Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet," *Proc. INFOCOM 2002*, vol. 2, pp. 1014–1023, 2002.

[6] J. 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, 1970.

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

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

[9] 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.

[10] ——, "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.

[11] C. Knessl, "On the sojourn time distribution in a finite capacity processor shared queue," *J. ACM*, vol. 40, no. 5, pp. 1238–1301, 1993.

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

[13] M. Harchol-Balter and A. B. Downey, "Exploiting process lifetime distributions for dynamic load balancing," *ACM Trans. Comput. Syst.*, vol. 15, no. 3, pp. 253–285, 1997.

[14] L. Kleinrock, *Queueing Systems. Volume 1: Theory.*. New York: Wiley, 1975.

[15] R. Litjens, J. V. D. Berg, and R. Boucherie, "Throughputs in processor sharing models for integrated stream and elastic traffic," *Memorandum Universiteit Twente, Faculteit der Toegepaste Wiskunde*, 2004, no. 1708.