J. Appl. Prob. **42,** 478–490 (2005)
Printed in Israel

PROCESSOR-SHARING AND RANDOM-SERVICE QUEUES WITH SEMI-MARKOVIAN ARRIVALS

DE-AN WU * AND
HIDEAKI TAKAGI,** *University of Tsukuba*

Abstract

We consider single-server queues with exponentially distributed service times, in which the arrival process is governed by a semi-Markov process (SMP). Two service disciplines, processor sharing (PS) and random service (RS), are investigated. We note that the sojourn time distribution of a type-l customer who, upon his arrival, meets k customers already present in the SMP/M/1/PS queue is identical to the waiting time distribution of a type-l customer who, upon his arrival, meets $k+1$ customers already present in the SMP/M/1/RS queue. Two sets of system equations, one for the joint transform of the sojourn time and queue size distributions in the SMP/M/1/PS queue, and the other for the joint transform of the waiting time and queue size distributions in the SMP/M/1/RS queue, are derived. Using these equations, the mean sojourn time in the SMP/M/1/PS queue and the mean waiting time in the SMP/M/1/RS queue are obtained. We also consider a special case of the SMP in which the interarrival time distribution is determined only by the type of the customer who has most recently arrived. Numerical examples are also presented.

Keywords: Queue; semi-Markov arrival process; processor sharing; random service; sojourn time; waiting time

2000 Mathematics Subject Classification: Primary 60K25
Secondary 60K15

1. Introduction

We study queueing systems with a single server, in which the arrival process is governed by a semi-Markov process (SMP). The service time follows an exponential distribution and the capacity of the waiting room is infinite. Two service disciplines are considered: (i) processor sharing (PS), where, when there are k customers in the system, each receives service at rate $1/k$; and (ii) random service (RS), where, when the server becomes available, the next customer to enter service is chosen at random among all waiting customers. The systems described above are denoted by SMP/M/1/PS and SMP/M/1/RS, respectively, throughout the paper.

The PS discipline is the limiting case of the round-robin discipline as the quantum of service time approaches 0, and allows for efficient and fair distribution of resources. PS queues have been widely used in modeling computer and communication systems. Since Coffman *et al.* [3]

Received 11 February 2003; revision received 27 August 2004.
* Postal address: Doctoral Program in Policy and Planning Sciences, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305-8573, Japan.
** Postal address: Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba-shi, Ibaraki 305-8573, Japan. Email address: takagi@sk.tsukuba.ac.jp

first analyzed an M/M/1/PS queue, several queueing systems with processor-sharing service have been studied, for example M/G/1/PS queues [14], [19], GI/M/1/PS queues [5], [10], [15], and GI/G/1/PS queues [16]. A survey of works on PS queues prior to 1987 was made by Yashkov [20], who cited many other references.

Early work on the PS queue [3] was motivated by the study of multiuser mainframe computer systems. Recently, PS queues have become an important tool for the performance evaluation of computer networks and web servers. Consider a number of independent transmissions on a network. Transmission Control Protocol (TCP) controls the transmission rate of a sender by adapting the congestion window size. Assuming that the TCP's feedback and control mechanism is perfect and absolutely fair, the bandwidth of a common bottleneck link will be shared equally among the active connections. This situation can be modeled as a PS queue [11]. Since the PS discipline allows shorter jobs to finish before longer jobs, most web servers employ PS-based algorithms to achieve the best 'user-perceived' performance in terms of fairness and response time.

On the other hand, correlated-input-process models are of increasing interest for the performance evaluation of computer networks. This is due to the fact that the superposition of video and other traffic sources, such as voice and data, may yield a complex arrival process characterized by high peak rate, general marginal distribution, and correlation between arrivals. The arrival process governed by an SMP can model this kind of autocorrelated traffic. We would like to mention that many other input processes, e.g. the special semi-Markov process (SSMP) [6], [18] and the two-state Markov-modulated Poisson process (MMPP(2)) [9], are special cases of the SMP. In [1], a queue with MMPP(2) arrivals and PS service discipline was used to model a web server, and the performance values of the web server were obtained by simulation. To the best of our knowledge, however, there have been no studies on queueing systems with SMP arrivals and PS discipline.

Ramaswami [15] found the first two moments of the sojourn time distribution in a GI/M/1/PS queue (he noted an error in [3]). Cohen [5] pointed out that the sojourn time distribution of a customer who, upon his arrival, meets k customers already present in a GI/M/1/PS queue is identical to the waiting time distribution of a customer who, upon his arrival, meets $k+1$ customers already present in a GI/M/1/RS queue. We note that the same relation exists between the SMP/M/1/PS queue and the SMP/M/1/RS queue. This is our motivation for analyzing the two queueing systems SMP/M/1/PS and SMP/M/1/RS together in this paper.

The rest of the paper is organized as follows. In Section 2, we define the semi-Markov arrival process and review some results about the queue size distribution before arrivals in SMP/M/1 queues. In Section 3, the sojourn time distribution of an arbitrary customer in the SMP/M/1/PS queue is considered. The waiting time distribution of an arbitrary customer in the SMP/M/1/RS queue is analyzed in Section 4. A special case of the SMP is investigated in Section 5. In Section 6, we give two numerical examples: one is the case of a general two-state semi-Markov arrival process and the other is the case of burst arrivals generated by an SMP.

2. Preliminaries

2.1. Semi-Markov arrival process

The semi-Markov arrival process can be described as follows [2]. There are L types of customer, numbered 1 through L. Customers arrive at time epochs $0 = T_0 < T_1 < T_2 < \cdots$, and $A_n := T_n - T_{n-1}$, $n > 1$, is the interarrival time, with $A_0 := 0$. Let $S^{(n)}$ denote the type of a customer arriving at epoch T_n. For a given sequence of arrival epochs, all interarrival times

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

are mutually independent. It is assumed that A_{n+1} and $S^{(n+1)}$ depend only on $S^{(n)}$, i.e.

$$\begin{aligned}&\mathrm{P}\{S^{(n+1)} = m,\ A_{n+1} \le t \mid S^{(0)}, \ldots, S^{(n)},\ A_1, \ldots, A_n\}\\ &\quad = \mathrm{P}\{S^{(n+1)} = m,\ A_{n+1} \le t \mid S^{(n)}\}, \qquad m = 1, \ldots, L,\ t \ge 0.\end{aligned}$$

Let

$$a_{lm}(t) := \mathrm{P}\{S^{(n+1)} = m,\ A_{n+1} \le t \mid S^{(n)} = l\}$$

be the probability that the arrival process moves from state l to state m in time t. We note that $a_{lm}(\infty)$ is the probability that the arrival of a type-l customer is followed by the arrival of a type-m customer. Let us define the Laplace–Stieltjes transform (LST) of $a_{lm}(t)$ as

$$\alpha_{lm}(s) := \int_0^\infty \mathrm{e}^{-st}\, \mathrm{d}a_{lm}(t).$$

We also define the matrix $\boldsymbol{A}(s) := \{\alpha_{lm}(s)\}$. It is noted that $\boldsymbol{A}(0) = \{a_{lm}(\infty)\}$ is a stochastic matrix; thus

$$\boldsymbol{A}(0)\mathbf{1} = \mathbf{1},$$

where $\mathbf{1} := [1, \ldots, 1]^\top$ ('$^\top$' denotes transpose and $\mathbf{1}$ is the identity vector with L elements). If $\boldsymbol{\pi} := [\pi_1, \ldots, \pi_L]$ is the stationary distribution of the stochastic matrix $\boldsymbol{A}(0)$, we have

$$\boldsymbol{\pi}\boldsymbol{A}(0) = \boldsymbol{\pi}, \qquad \boldsymbol{\pi}\mathbf{1} = 1.$$

Without much loss of generality, we assume that the Markov chain $\{S^{(n)},\ n = 0, 1, 2, \ldots\}$ is ergodic. For a real number $s \ge 0$, if $\lambda_{\mathrm{M}}(s)$ denotes the eigenvalue of the matrix $\boldsymbol{A}(s)$ with the maximum absolute value, then [2, Equation (9)]

$$\alpha = -\frac{\mathrm{d}}{\mathrm{d}s}\lambda_{\mathrm{M}}(s)\bigg|_{s=0+}$$

is the mean interarrival time, defined by

$$\alpha := \sum_{l=1}^{L} \pi_l \sum_{m=1}^{L} \int_0^\infty t\, \mathrm{d}a_{lm}(t).$$

2.2. Queue size distribution immediately before arrivals in SMP/M/1 queues

As in GI/M/1 queues [15], owing to the memoryless property of exponentially distributed service times, the queue size distribution immediately before arrivals in the SMP/M/1/PS queue is the same as that in the corresponding SMP/M/1 queue with first-in–first-out (FIFO) service discipline. Here we review some results about the SMP/M/1/FIFO queue from [2], which will be used in studying the sojourn time distribution in the SMP/M/1/PS queue. Let μ be the service rate.

The following theorem is a special case of the result of [2, p. 366], which is fundamental to the analysis of SMP/M/1 queues.

Theorem 1. *The equation*

$$\det[z\boldsymbol{I} - \boldsymbol{A}(s + \mu - \mu z)] = 0 \tag{1}$$

has exactly L solutions, $z = \gamma_1(s), \gamma_2(s), \ldots, \gamma_L(s)$, within the unit circle $|z| = 1$ if $\mathrm{Re}(s) > 0$. Here, $\boldsymbol{I}$ denotes the $L \times L$ identity matrix.

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

This theorem is the matrix version of Lemma 1 of Takács [17, p. 113] for a GI/M/1 queue, and it can be proved by application of permutation theory and Rouché's theorem [12], [13]. We denote the distinct solutions of (1) by $\gamma^{(1)}(s), \gamma^{(2)}(s), \ldots, \gamma^{(M)}(s)$, with $M \leq L$.

For the analysis of SMP/M/1 queues, we impose the following assumption, which is the same as the one in [2].

Assumption 1. *All the elementary divisors [7, p. 142] of the matrix* $\boldsymbol{A}(s + \mu - \mu\gamma^{(i)}(s))$ *corresponding to the eigenvalue* $\gamma^{(i)}(s)$ *are of the first degree, for* $i = 1, \ldots, M$.

We note that the multiple eigenvalues are not ruled out. The matrix $\boldsymbol{A}(s + \mu - \mu\gamma^{(i)}(s))$ may have elementary divisors not of the first degree if they correspond to eigenvalues other than $\gamma^{(i)}(s)$. If we denote by $\boldsymbol{g}_i(s)$ the left-eigenvector corresponding to the eigenvalue $\gamma_i(s)$, we have

$$\boldsymbol{g}_i(s)\{\gamma_i(s)\boldsymbol{I} - \boldsymbol{A}(s + \mu - \mu\gamma_i(s))\} = \boldsymbol{0}, \qquad i = 1, \ldots, L,$$

where $\boldsymbol{0} := [0, \ldots, 0]$ is the zero vector with L elements. We collect all the eigenvectors corresponding to the eigenvalues $\gamma_1(s), \ldots, \gamma_L(s)$ into an $L \times L$ matrix $\boldsymbol{G}(s)$, as follows:

$$\boldsymbol{G}(s) := [\boldsymbol{g}_1(s), \boldsymbol{g}_2(s), \ldots, \boldsymbol{g}_L(s)]^\top.$$

Moreover, by $\boldsymbol{\Gamma}(s)$ we denote an $L \times L$ diagonal matrix with elements $\gamma_1(s), \gamma_2(s), \ldots, \gamma_L(s)$.

We next consider the Markov chain $\{(X^{(n)}, S^{(n)}),\ n = 0, 1, 2, \ldots\}$, where $X^{(n)}$ denotes the number of customers seen by the nth SMP arrival. The following result is cited from Theorem 5 of [2], which gives the queue size distribution immediately before arrivals in the SMP/M/1/FIFO queue. Thus, it is also the queue size distribution immediately before arrivals in the SMP/M/1/PS queue.

Theorem 2. *Under Assumption 1, all states of the Markov chain* $\{(X^{(n)}, S^{(n)}),\ n = 0, 1, 2, \ldots\}$ *are ergodic if* $\alpha\mu > 1$. *In this case,*

$$\varpi_k^{(l)} := \lim_{n\to\infty} \mathrm{P}\{X^{(n)} = k,\ S^{(n)} = l\}, \qquad k = 0, 1, 2, \ldots,\ l = 1, 2, \ldots, L,$$

exist and are independent of the initial distribution. Letting $\boldsymbol{\varpi}_k := [\varpi_k^{(1)}, \varpi_k^{(2)}, \ldots, \varpi_k^{(L)}]$, *we have*

$$\boldsymbol{\varpi}_k = \boldsymbol{\pi}\boldsymbol{G}^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma})\boldsymbol{\Gamma}^k\boldsymbol{G}, \qquad k = 0, 1, 2, \ldots, \tag{2}$$

where $\boldsymbol{\Gamma} := \boldsymbol{\Gamma}(0)$ *and* $\boldsymbol{G} := \boldsymbol{G}(0)$.

We rewrite (2) in scalar form as

$$\varpi_k^{(m)} = \sum_{l=1}^{L} \beta_l \gamma_l^k g_{lm}, \qquad k = 0, 1, 2, \ldots,\ m = 1, 2, \ldots, L, \tag{3}$$

where β_l is the lth element of the row vector

$$\boldsymbol{\beta} := \boldsymbol{\pi}\boldsymbol{G}^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma}), \tag{4}$$

$\gamma_l := \gamma_l(0)$, and g_{lm} is the (l, m)th element of the matrix $\boldsymbol{G}$.

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

3. Sojourn time in the SMP/M/1/PS queue

We now derive the sojourn time distribution in the SMP/M/1/PS queue. Let us focus on a *tagged* customer of type l who finds k other customers in the system upon his arrival. Let $S_k^{(l)}(t)$ denote the sojourn time distribution of this tagged customer. We define

$$A_{lm}(j,t) := \int_0^t \frac{(\mu x)^j}{j!} \mathrm{e}^{-\mu x}\, \mathrm{d}a_{lm}(x)$$

as the probability that exactly j customers are served in the time between the arrival of a type-l customer and the immediately following arrival of a type-m customer, when this interarrival time is less than t.

Lemma 1. *The functions $S_k^{(l)}(t)$ satisfy the following relations, where $A_{lm}(j,t) * S_k^{(m)}(t)$ denotes the convolution of $A_{lm}(j,t)$ and $S_k^{(m)}(t)$:*

$$\begin{aligned} S_k^{(l)}(t) = {} & \sum_{m=1}^{L} \sum_{j=1}^{k+1} \frac{1}{k+1} \int_0^t [a_{lm}(\infty) - a_{lm}(x)] \frac{\mu(\mu x)^{j-1}}{(j-1)!} \mathrm{e}^{-\mu x}\, \mathrm{d}x \\ & + \sum_{m=1}^{L} \sum_{j=0}^{k} \frac{k+1-j}{k+1} A_{lm}(j,t) * S_{k+1-j}^{(m)}(t), \qquad l = 1, 2, \ldots, L,\ k = 0, 1, 2, \ldots. \end{aligned} \tag{5}$$

Proof. Our proof extends the method of [15]. We write

$$S_k^{(l)}(t) = F_k^{(l)}(t) + B_k^{(l)}(t), \tag{6}$$

where $F_k^{(l)}(t)$ is the probability that the tagged customer, being of type l and finding k other customers present, ends his service *before* the next arrival and has a sojourn time less than t, and $B_k^{(l)}(t)$ is the probability that the tagged customer, being of type l and finding k other customers present, ends his service *after* the next arrival and has a sojourn time less than t.

Consider $F_k^{(l)}(t)$. Owing to the memoryless property of the exponentially distributed service time, all customers present at time x have the same distribution for the residual sojourn time. If there is a departure in a short time interval $(x, x + \Delta x]$, each customer present at time x has the same chance to depart. Thus, if at least j customers end their services before the next arrival, then the probability that the tagged customer is the jth to leave the system is given by

$$\frac{1}{j} \binom{k}{j-1} \binom{k+1}{j}^{-1} = \frac{1}{k+1}.$$

Conditioning on both the type of the next arrival and the number of departures before the next arrival, we have

$$F_k^{(l)}(t) = \sum_{m=1}^{L} \sum_{j=1}^{k+1} \frac{1}{k+1} \int_0^t [a_{lm}(\infty) - a_{lm}(x)] \frac{\mu(\mu x)^{j-1}}{(j-1)!} \mathrm{e}^{-\mu x}\, \mathrm{d}x. \tag{7}$$

Now consider $B_k^{(l)}(t)$. The probability that the tagged customer is not one of the j customers who depart from the system before the next arrival is given by

$$\binom{k}{j} \binom{k+1}{j}^{-1} = \frac{k+1-j}{k+1}.$$

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

Conditioning on the length of the interarrival time, the type of the next arrival, and the number of departures before the next arrival, we obtain

$$B_k^{(l)}(t) = \sum_{m=1}^{L} \sum_{j=0}^{k} \frac{k+1-j}{k+1} A_{lm}(j,t) * S_{k+1-j}^{(m)}(t). \tag{8}$$

Substituting (7) and (8) into (6) gives (5).

We remark that if there is only one type of customer, then our Lemma 1 reduces to Lemma 1 of [15].

Let us define the generating function of the LST of $S_k^{(l)}(t)$ as

$$\sigma^{(l)}(z,s) := \sum_{k=0}^{\infty} \sigma_k^{(l)}(s) z^k, \qquad l = 1, 2, \ldots, L,$$

where

$$\sigma_k^{(l)}(s) := \int_0^{\infty} \mathrm{e}^{-st}\, \mathrm{d}S_k^{(l)}(t), \qquad k = 0, 1, 2, \ldots.$$

Introducing the column vector $\boldsymbol{\sigma}(z,s) := [\sigma^{(1)}(z,s), \sigma^{(2)}(z,s), \ldots, \sigma^{(L)}(z,s)]^{\top}$, we have the following theorem.

Theorem 3. *The vector $\boldsymbol{\sigma}(z,s)$ satisfies the differential equation*

$$[z\boldsymbol{I} - \boldsymbol{A}(s+\mu-\mu z)] \frac{\partial \boldsymbol{\sigma}(z,s)}{\partial z} + \boldsymbol{\sigma}(z,s) = \frac{\mu}{(1-z)(s+\mu-\mu z)} [\boldsymbol{A}(0) - \boldsymbol{A}(s+\mu-\mu z)]\mathbf{1}. \tag{9}$$

Proof. Let us introduce the notation

$$\psi^{(l)}(z,s) := \sum_{k=0}^{\infty} (k+1)\sigma_k^{(l)}(s) z^k.$$

It is easy to verify that

$$\psi^{(l)}(z,s) = \sigma^{(l)}(z,s) + z \frac{\partial \sigma^{(l)}(z,s)}{\partial z}. \tag{10}$$

By taking the LST of (5), multiplying by $(k+1)z^k$, and summing over $k = 0, 1, 2, \ldots$, we obtain

$$\psi^{(l)}(z,s) = \frac{\mu}{(1-z)(s+\mu-\mu z)} \sum_{m=1}^{L} [\alpha_{lm}(0) - \alpha_{lm}(s+\mu-\mu z)] + \frac{1}{z} \sum_{m=1}^{L} \alpha_{lm}(s+\mu-\mu z)[\psi^{(m)}(z,s) - \sigma^{(m)}(z,s)]. \tag{11}$$

Substituting (10) into (11) yields

$$z \frac{\partial \sigma^{(l)}(z,s)}{\partial z} + \sigma^{(l)}(z,s) = \frac{\mu}{(1-z)(s+\mu-\mu z)} \sum_{m=1}^{L} [\alpha_{lm}(0) - \alpha_{lm}(s+\mu-\mu z)] + \sum_{m=1}^{L} \alpha_{lm}(s+\mu-\mu z) \frac{\partial \sigma^{(m)}(z,s)}{\partial z}, \tag{12}$$

and rewriting (12) in matrix form gives (9).

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

Remark 1. Note that (9) has L singularities at $z = \gamma_l(s), l = 1, \ldots, L$. This means that there may be solutions to (9) that are not analytic at $z = \gamma_l(s)$; it seems difficult to find a solution to the differential equations in (9) in a general case. Theoretically speaking, (9) can be solved in the real domain $z \in (-1, 1)$, with $z \neq \gamma_l(s)$, by using the multiplicative integral [8, p. 132]. However, it seems difficult to obtain $\boldsymbol{\sigma}(z, s)$ explicitly.

Let $\sigma(s)$ denote the LST of the sojourn time of an arbitrary customer. It follows that

$$\sigma(s) = \sum_{m=1}^{L} \sum_{k=0}^{\infty} \varpi_k^{(m)} \sigma_k^{(m)}(s). \tag{13}$$

Substituting (3) into (13) yields

$$\begin{aligned}\sigma(s) &= \sum_{m=1}^{L} \sum_{k=0}^{\infty} \sum_{l=1}^{L} \beta_l g_{lm} \gamma_l^k \sigma_k^{(m)}(s) = \sum_{l=1}^{L} \sum_{m=1}^{L} \beta_l g_{lm} \sigma^{(m)}(\gamma_l, s) \\ &= \sum_{l=1}^{L} \beta_l \boldsymbol{g}_l \boldsymbol{\sigma}(\gamma_l, s),\end{aligned}$$

where $\boldsymbol{g}_l$ is the lth row of the matrix $\boldsymbol{G}$. Thus, the mean sojourn time σ of an arbitrary customer is given by

$$\sigma = -\sum_{l=1}^{L} \beta_l \boldsymbol{g}_l \frac{\partial}{\partial s} \boldsymbol{\sigma}(\gamma_l, s)\bigg|_{s=0+}. \tag{14}$$

Theorem 4. *The mean sojourn time of an arbitrary customer in the SMP/M/1/PS queue is given by*

$$\sigma = \frac{1}{\mu} \boldsymbol{\pi} \boldsymbol{G}^{-1} (\boldsymbol{I} - \boldsymbol{\Gamma})^{-1} \boldsymbol{G} \mathbf{1}. \tag{15}$$

Proof. Let us introduce the column vector

$$\boldsymbol{v}_l(s) := \frac{\partial}{\partial z} \boldsymbol{\sigma}(z, s)\bigg|_{z=\gamma_l}, \qquad l = 1, 2, \ldots, L.$$

Evaluating both sides of (9) at $z = \gamma_l$ yields

$$[\gamma_l \boldsymbol{I} - \boldsymbol{A}(s + \mu - \mu\gamma_l)] \boldsymbol{v}_l(s) + \boldsymbol{\sigma}(\gamma_l, s) = \frac{\mu}{(1 - \gamma_l)(s + \mu - \mu\gamma_l)} [\boldsymbol{A}(0) - \boldsymbol{A}(s + \mu - \mu\gamma_l)] \mathbf{1}. \tag{16}$$

Differentiating (16) with respect to s and taking the limit as s approaches $0+$ gives

$$[\gamma_l \boldsymbol{I} - \boldsymbol{A}(\mu - \mu\gamma_l)] \boldsymbol{v}_l'(0) + \frac{\partial}{\partial s} \boldsymbol{\sigma}(\gamma_l, s)\bigg|_{s=0+} = -\frac{[\boldsymbol{A}(0) - \boldsymbol{A}(\mu - \mu\gamma_l)] \mathbf{1}}{\mu (1 - \gamma_l)^3}, \tag{17}$$

where we have used

$$\boldsymbol{v}_l(0) = \frac{\partial}{\partial z} \boldsymbol{\sigma}(z, 0)\bigg|_{z=\gamma_l} = \frac{\mathrm{d}}{\mathrm{d}z} \left(\frac{1}{1 - z} \right)\bigg|_{z=\gamma_l} \mathbf{1} = \frac{1}{(1 - \gamma_l)^2} \mathbf{1}.$$

Recall that $\boldsymbol{g}_l$ is the left-eigenvector of the matrix $\boldsymbol{A}(\mu - \mu\gamma_l)$ corresponding to the eigenvalue γ_l. It follows that

$$\boldsymbol{g}_l [\gamma_l \boldsymbol{I} - \boldsymbol{A}(\mu - \mu\gamma_l)] = \mathbf{0}.$$

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

Multiplying (17) on the left by $\boldsymbol{g}_l$ gives

$$-\boldsymbol{g}_l \frac{\partial}{\partial s}\boldsymbol{\sigma}(\gamma_l, s)\bigg|_{s=+0} = \frac{\boldsymbol{g}_l \mathbf{1}}{\mu(1-\gamma_l)^2}. \tag{18}$$

Using (4) and (18) in (14) gives (15).

4. Waiting time in the SMP/M/1/RS queue

We proceed to analyze the SMP/M/1/RS queue. In this system, the service time follows an exponential distribution with rate μ, the arrival process is governed by the semi-Markov process described in Section 2, and the service discipline is random. The random service discipline is described as follows: at the end of a service, the next customer to be served is selected at random among all the customers present in the queue. Since the order of service does not influence the unfinished work of the system, the queue size distribution is independent of the order of service. Hence, the queue size distribution immediately before arrivals in the SMP/M/1/RS queue is identical to that in the corresponding SMP/M/1/FIFO queue.

Let $W_k^{(l)}(t)$ denote the waiting time distribution of a *tagged* customer of type l who finds $k+1$ other customers in the system upon his arrival. We then have the following lemma.

Lemma 2. *The functions $W_k^{(l)}(t)$ satisfy the relations in (5), with $S_k^{(l)}(t)$ replaced by $W_k^{(l)}(t)$, for $l = 1, 2, \ldots, L$ and $k = 0, 1, 2, \ldots$.*

Proof. The proof is similar to that of Lemma 1.

Remark 2. Comparing Lemma 1 and Lemma 2, we note that the sojourn time distribution of a type-l customer who, upon his arrival, meets k customers already present in the SMP/M/1/PS queue is identical to the waiting time distribution of a type-l customer who, upon his arrival, meets $k+1$ customers already present in the SMP/M/1/RS queue. This is an extension of the relation between the GI/M/1/PS queue and the GI/M/1/RS queue mentioned by Cohen [5].

Let us define the generating function of the LST of $W_k^{(l)}(t)$ as

$$w^{(l)}(z, s) := \sum_{k=0}^{\infty} w_k^{(l)}(s) z^k,$$

where

$$w_k^{(l)}(s) := \int_0^{\infty} \mathrm{e}^{-st}\, \mathrm{d}W_k^{(l)}(t), \qquad k = 0, 1, 2, \ldots.$$

By the method used previously to derive (9), we obtain the following result.

Theorem 5. *The vector $\boldsymbol{w}(z, s) := [w^{(1)}(z, s), w^{(2)}(z, s), \ldots, w^{(L)}(z, s)]^{\top}$ satisfies the differential equation in (9), with $\boldsymbol{\sigma}(z, s)$ replaced by $\boldsymbol{w}(z, s)$.*

Let $w(s)$ denote the LST of the waiting time distribution of an arbitrary customer. Conditioning on both the number of customers present in the system immediately before the arrival of the arbitrary customer and the type of the arbitrary customer, we obtain

$$w(s) = \sum_{m=0}^{L} \varpi_0^{(m)} + \sum_{m=1}^{L} \sum_{k=1}^{\infty} \varpi_k^{(m)} w_{k-1}^{(m)}(s). \tag{19}$$

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

Using (3) in (19) gives

$$\begin{aligned} w(s) &= \sum_{m=1}^{L}\sum_{l=1}^{L}\beta_l g_{lm} + \sum_{m=1}^{L}\sum_{k=0}^{\infty}\sum_{l=1}^{L}\beta_l g_{lm}\gamma_l^{k+1} w_k^{(m)}(s) \\ &= \sum_{l=1}^{L}\sum_{m=1}^{L}\beta_l g_{lm} + \sum_{l=1}^{L}\sum_{m=1}^{L}\beta_l g_{lm}\gamma_l w^{(m)}(\gamma_l, s) \\ &= \sum_{l=1}^{L}\beta_l \boldsymbol{g}_l \mathbf{1} + \sum_{l=1}^{L}\beta_l\gamma_l \boldsymbol{g}_l \boldsymbol{w}(\gamma_l, s). \end{aligned}$$

Therefore, the mean waiting time w of an arbitrary customer is given by

$$w = -\sum_{l=1}^{L}\beta_l\gamma_l \boldsymbol{g}_l \frac{\partial}{\partial s}\boldsymbol{w}(\gamma_l, s)\bigg|_{s=0+}. \tag{20}$$

Theorem 6. *The mean waiting time of an arbitrary customer in the SMP/M/1/RS queue is given by*

$$w = \frac{1}{\mu}\boldsymbol{\pi}\boldsymbol{G}^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}\boldsymbol{G}\mathbf{1}. \tag{21}$$

Proof. By the method used to derive (18), we obtain

$$-\boldsymbol{g}_l \frac{\partial}{\partial s}\boldsymbol{w}(\gamma_l, s)\bigg|_{s=0+} = \frac{\boldsymbol{g}_l\mathbf{1}}{\mu(1-\gamma_l)^2}, \qquad l = 1, 2, \ldots, L. \tag{22}$$

Using (4) and (22) in (20) gives (21).

5. Special semi-Markovian arrival process

In this section, we consider the sojourn time in the SMP/M/1/PS queue and the waiting time in the SMP/M/1/RS queue in a special case of the semi-Markov arrival process. That is, we assume that the interarrival time distribution is determined only by the type of the immediately prior arrival; we write this as $\boldsymbol{A}(s) = [\alpha_1(s), \alpha_2(s), \ldots, \alpha_L(s)]^\top \mathbf{1}^\top$, where

$$\alpha_l(s) := \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}\,\mathrm{P}\{A_{n+1} \le t \mid S^{(n)} = l\}, \qquad l = 1, 2, \ldots, L.$$

It is then easy to verify that

$$\det[z\boldsymbol{I} - \boldsymbol{A}(s)] = z^{L-1}[z - \alpha(s)]$$

and that $z[z - \alpha(s)]$ is the minimal polynomial [7, p. 89] of $\boldsymbol{A}(s)$, where

$$\alpha(s) := \sum_{l=1}^{L}\alpha_l(s).$$

Let us first consider the sojourn time in the SMP/M/1/PS queue for this special semi-Markov arrival process. If Assumption 1 is satisfied, then the solutions to the equation

$$\det[z\boldsymbol{I} - \boldsymbol{A}(s + \mu - \mu z)] = 0$$

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

are $\gamma_1(s) = \gamma_2(s) = \cdots = \gamma_{L-1}(s) = 0$ and $\gamma_L(s) = \gamma(s)$, where $\gamma(s)$ is the solution to the equation $z - \alpha(s + \mu - \mu z) = 0$ within the unit circle $|z| = 1$; hereafter, we write γ for $\gamma(0)$. Therefore, $\boldsymbol{\Gamma}$ becomes a diagonal matrix with the elements $0, 0, \ldots, 0, \gamma$. We note that the left-eigenvector of $\boldsymbol{A}(\mu - \mu\gamma)$ corresponding to the eigenvalue γ is $\mathbf{1}^\top$, which is the last row of the matrix $\boldsymbol{G}$. The last column of the matrix $\boldsymbol{G}^{-1}$ is

$$\bar{\boldsymbol{g}}_L = \frac{1}{\gamma}[\alpha_1(\mu - \mu\gamma), \alpha_2(\mu - \mu\gamma), \ldots, \alpha_L(\mu - \mu\gamma)]^\top,$$

which is the right-eigenvector corresponding to γ. It follows that

$$\begin{aligned}
&\boldsymbol{G}^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma})^{-1}\boldsymbol{G}\mathbf{1} \qquad (23)\\
&= \begin{bmatrix} \cdot & \cdot & \cdots & \alpha_1(\mu - \mu\gamma)/\gamma \\ \cdot & \cdot & \cdots & \alpha_2(\mu - \mu\gamma)/\gamma \\ \vdots & \vdots & \ddots & \vdots \\ \cdot & \cdot & \cdots & \alpha_L(\mu - \mu\gamma)/\gamma \end{bmatrix} \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/(1-\gamma) \end{bmatrix} \\
&\quad \times \begin{bmatrix} \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 + \dfrac{\alpha_1(\mu - \mu\gamma)}{1-\gamma} & \dfrac{\alpha_1(\mu - \mu\gamma)}{1-\gamma} & \cdots & \dfrac{\alpha_1(\mu - \mu\gamma)}{1-\gamma} \\ \dfrac{\alpha_2(\mu - \mu\gamma)}{1-\gamma} & 1 + \dfrac{\alpha_2(\mu - \mu\gamma)}{1-\gamma} & \cdots & \dfrac{\alpha_2(\mu - \mu\gamma)}{1-\gamma} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\alpha_L(\mu - \mu\gamma)}{1-\gamma} & \dfrac{\alpha_L(\mu - \mu\gamma)}{1-\gamma} & \cdots & 1 + \dfrac{\alpha_L(\mu - \mu\gamma)}{1-\gamma} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \\
&= \mathbf{1} + \frac{\gamma L}{1-\gamma}\bar{\boldsymbol{g}}_L, \qquad (24)
\end{aligned}$$

where we have used $\boldsymbol{G}^{-1}\boldsymbol{G} = \boldsymbol{I}$. Substituting (24) into (15) gives

$$\sigma = \frac{1}{\mu(1-\gamma)}. \qquad (25)$$

We remark that if the number of types of customer is one for which the LST of the interarrival time distribution is $\alpha(s)$, then (25) reduces to the mean sojourn time in a GI/M/1/PS queue, given by [15, p. 440, Equation (8)].

We also consider the waiting time in the SMP/M/1/RS queue for the special semi-Markov arrival process described above. In the same way as we derived (24), we obtain

$$\boldsymbol{G}^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}\boldsymbol{G}\mathbf{1} = \frac{\gamma}{1-\gamma}\mathbf{1}. \qquad (26)$$

Using (26) in (21) yields

$$w = \frac{\gamma}{\mu(1-\gamma)}. \qquad (27)$$

Hence, if there is a single type of customer for which the LST of the interarrival time distribution is $\alpha(s)$, then (27) is reduced to the mean waiting time in the GI/M/1/RS queue. This queue is treated in Cohen [4, p. 443], but he does not comment on this reduction in [5].

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

6. Numerical examples

In this section, we illustrate the results for the SMP/M/1/PS queues obtained in the previous sections in two examples: one for a general two-state SMP and the other for an SMP that can model a burst arrival process. In both examples, we assume that

$$\boldsymbol{A}(s) = \begin{bmatrix} (1-p)b_{11}(s) & pb_{12}(s) \\ qb_{21}(s) & (1-q)b_{22}(s) \end{bmatrix}, \qquad p, q \in [0,1], \tag{28}$$

where $b_{lm} := \int_0^\infty \mathrm{e}^{-st}\,\mathrm{d}\,\mathrm{P}\{A_{n+1} \le t \mid S^{(n)} = l,\ S^{(n+1)} = m\},\ l, m = 1, 2$. It can be shown that the overall arrival rate is given by

$$\begin{aligned} \lambda &= [\pi_1, \pi_2][(1-p)\beta_{11} + p\beta_{12}\ ,\ q\beta_{21} + (1-q)\beta_{22}]^\top \\ &= \frac{q\beta_{11} + p\beta_{22} + pq(-\beta_{11} - \beta_{22} + \beta_{12} + \beta_{21})}{p+q}, \end{aligned} \tag{29}$$

where $\beta_{lm} = -1/b'_{lm}(0),\ l, m = 1, 2$, and

$$[\pi_1, \pi_2] = \left[\frac{q}{p+q}, \frac{p}{p+q}\right]$$

is the stationary distribution of the stochastic matrix $\boldsymbol{A}(0)$. We also assume that the service time is exponentially distributed with rate $\mu = 10$.

Example 1. Two SMP/M/1/PS queues are considered. The first queue is denoted by SMP(M)/M/1/PS, and we take

$$b_{11}(s) = b_{22}(s) = \frac{\lambda_1}{s+\lambda_1} \quad \text{and} \quad b_{12}(s) = b_{21}(s) = \frac{\lambda_2}{s+\lambda_2},$$

i.e. the interarrival time in the SMP is exponentially distributed. The second queue is denoted by SMP(Er)/M/1/PS, with

$$b_{11}(s) = b_{22}(s) = \left(\frac{2\lambda_1}{s+2\lambda_1}\right)^2 \quad \text{and} \quad b_{12}(s) = b_{21}(s) = \left(\frac{2\lambda_2}{s+2\lambda_2}\right)^2,$$

i.e. the interarrival time in the SMP follows an Erlang(2) distribution. We can now calculate the mean sojourn time of an arbitrary customer. For this purpose, the following parameters are used: $p = 0.2, q = 0.3$, and $\lambda_2 = \frac{1}{2}\lambda_1$.

In Figure 1, we plot the performance values as functions of λ, defined in (29). For comparison, the mean sojourn times in an M/M/1/PS queue and an Er/M/1/PS queue, with the same arrival rates as in the SMP/M/1/PS queues, are also plotted. It is observed that the SMP(M) customers always receive worse treatment, i.e. have a longer mean sojourn time, than the Poisson customers; the same relation exists between the SMP(Er) and Erlang customers. Furthermore, the mean sojourn time of the SMP(M) customers is longer than that of the SMP(Er) customers. This is similar to the relation between the sojourn time of a Poisson customer in an M/M/1/PS queue and the sojourn time of an Erlang customer in an Er/M/1/PS queue.

Example 2. The MMPP(2) is often used to model bursty traffic on communication networks. The MMPP(2) is a stochastic process with two arrival states, where the arrival process in state l is a Poisson process with rate $\theta_l, l = 1, 2$. The time intervals during which the process stays in

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

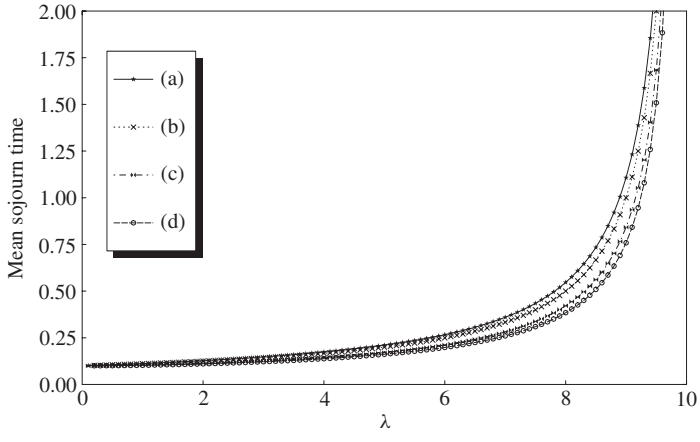

FIGURE 1: Sojourn times in the PS queues, where (a), (b), (c), and (d) represent the SMP(M)/M/1/PS, M/M/1/PS, SMP(Er)/M/1/PS, and Er/M/1/PS queues, respectively.

all states are exponentially distributed, with means $1/r_1$ and $1/r_2$ in states 1 and 2, respectively. The overall arrival rate in the MMPP(2) is given by [9]

$$\lambda = \frac{\theta_1 r_2 + \theta_2 r_1}{r_1 + r_2}. \tag{30}$$

It was shown in [6] that the point process generated by an MMPP(2) is stochastically equivalent to its matched two-state SMP, for which the LST matrix of the interarrival time distributions is given in (28), with

$$b_{11}(s) = b_{12}(s) = \frac{\lambda_1}{s + \lambda_1} \quad \text{and} \quad b_{21}(s) = b_{22}(s) = \frac{\lambda_2}{s + \lambda_2}. \tag{31}$$

For a given set of parameters $\{\theta_1, \theta_2, r_1, r_2\}$ of the MMPP(2), the parameters $\{p, q, \lambda_1, \lambda_2\}$ of the corresponding SMP are determined by the following set of equations:

$$\theta_1 + \theta_1 + r_1 + r_2 = \lambda_1 + \lambda_2, \tag{32}$$

$$r_1 + r_2 = q\lambda_2 + p\lambda_1, \tag{33}$$

$$\theta_1\theta_2 + \theta_1 r_2 + \theta_2 r_1 = \lambda_1\lambda_2, \tag{34}$$

$$\theta_1 r_2 + \theta_2 r_1 = (p + q)\lambda_1\lambda_2. \tag{35}$$

We first choose the parameters of an MMPP(2) that simulates a burst arrival process, and then construct an equivalent SMP(M) using (28), (31), and (32)–(35). The mean sojourn time of an arbitrary customer can then be determined from our analysis. Given the overall arrival rate λ of the SMP(M), we set $\theta_1 = 0.75\lambda$, $r_1 = 0.05$, and $r_2 = 0.95$. From (30), we then have $\theta_2 = 5.75\lambda$. The high arrival rate θ_2 implies that bursts of arrivals occur in 5% of the time interval in the regeneration cycle in this MMPP(2).

In Figure 2, we plot the performance values as functions of λ. The mean sojourn times in an SMP(Er)/M/1/PS are also plotted; here, each interarrival time follows the Erlang(2) distribution with the mean equal to the corresponding one in the SMP(M). For comparison, the performance values of an M/M/1/PS and an Er/M/1/PS are also plotted. We observe that the burst arrivals strongly influence the mean sojourn time in the SMP(M)/M/1/PS and SMP(Er)/M/1/PS queues.

https://doi.org/10.1239/jap/1118777183 Published online by Cambridge University Press

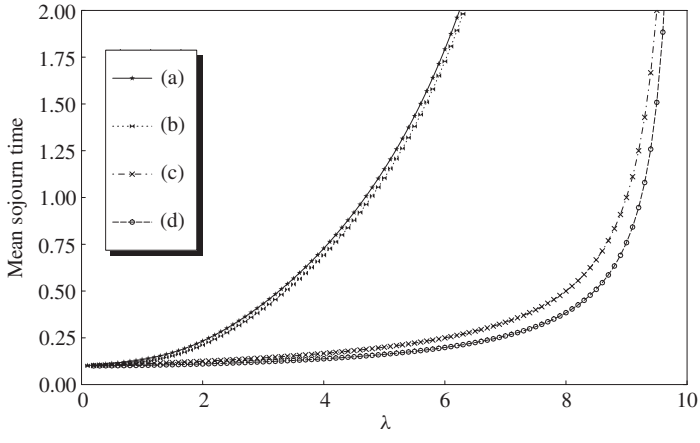

FIGURE 2: Sojourn times in the PS queues with burst arrivals, where (a), (b), (c), and (d) represent the SMP(M)/M/1/PS, SMP(Er)/M/1/PS, M/M/1/PS, and Er/M/1/PS queues, respectively.

References

[1] ANDERSSON, M., CAO, J., KIHL, M. AND NYBERG, C. (2003). Performance modeling of an Apache web server with bursty arrival traffic. In *Proc. Internat. Conf. Internet Comput.* (Las Vegas, June 2003), Vol. 2, pp. 508–514.

[2] ÇINLAR, E. (1967). Queues with semi-Markovian arrivals. *J Appl. Prob.* **4,** 365–379.

[3] COFFMAN, E. G., JR., MUNTZ, R. R. AND TROTTER, H. (1970). Waiting time distributions for processor-sharing systems. *J. Assoc. Comput. Mach.* **17,** 123–130.

[4] COHEN, J. W. (1982). *The Single Server Queue* (North-Holland Ser. Appl. Math. Mech. **8**), 2nd edn. North-Holland, Amsterdam.

[5] COHEN, J. W. (1984). On processor sharing and random service. *J. Appl. Prob.* **21,** 937.

[6] DING, W. (1991). A unified correlated input process model for telecommunication networks. In *Teletraffic and Datatraffic in a Period of Change*, eds A. Jensen and V. B. Iversen, Elsevier, Amsterdam, pp. 539–544.

[7] GANTMACHER, F. R. (2000). *The Theory of Matrices*, Vol. 1. AMS, Providence, RI.

[8] GANTMACHER, F. R. (2000). *The Theory of Matrices*, Vol. 2. AMS, Providence, RI.

[9] HEFFES, H. (1980). A class of data traffic processes–covariance function characterization and related queuing results. *Bell System Tech. J.* **59,** 897–929.

[10] JAGERMAN, D. AND SENGUPTA, B. (1991). The GI/M/1 processor-sharing queue and its heavy traffic analysis. *Commun. Statist. Stoch. Models* **7,** 379–395.

[11] KHERANI, A. AND KUMAR, A. (2002). Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet. In *Proc. IEEE INFOCOM* (New York, July 2002), Vol. 2, pp. 1014–1023.

[12] LOYNES, R. M. (1962). Stationary waiting-time distribution for single server queue. *Ann Math. Statist.* **33,** 1323–1339.

[13] NEUTS, M. F. (1966). The single server queue with Poisson input and semi-Markov service times. *J. Appl. Prob.* **3,** 202–230.

[14] OTT, T. J. (1984). The sojourn time distribution in the M/G/1 queue with processor sharing *J. Appl. Prob.* **21,** 360–378.

[15] RAMASWAMI, V. (1984). The sojourn time in the GI/M/1 queue with processor sharing. *J. Appl. Prob.* **21,** 437–442.

[16] SENGUPTA, B. (1991). An approximation for the sojourn-time distribution for the GI/G/1 processor sharing discipline queue. *Commun. Statist. Stoch. Models* **8,** 379–395.

[17] TAKÁCS, L. (1962). *Introduction to the Theory of Queues*. Oxford University Press.

[18] YAGYU, S. AND TAKAGI, H. (2002). A queueing model with input of MPEG fame sequence and interfering traffic. *J. Operat. Res. Soc. Japan* **45,** 317–338.

[19] YASHKOV, S. F. (1983). A derivation of response time distribution for an M/G/1 processor-sharing queue. *Problems Control Inf. Theory* **12,** 133–148.

[20] YASHKOV, S. F. (1987). Processor-sharing queues: some progress in analysis. *Queueing Systems* **2,** 1–17.