Информационные процессы, Том 6, № 2, 2006, стр. 160–163.

ON RANDOM ORDER OF SERVICE AND PROCESSOR SHARING [1]

S.F.Yashkov

Institute for Information Transmission Problems, 19, Bolshoi Karetny lane, 127994 Moscow GSP–4, Russia. E-mail: yashkov@iitp.ru

Received May 19, 2006

Abstract—We show that the stationary sojourn time distribution in the M/M/1 (egalitarian) processor sharing queue is closely related to the stationary waiting time distribution in the same queue under the random order of service. This enables us to obtain new results concerning tail behaviour of the sojourn time in the processor sharing queue, in particular, for the normalized maxima of the sojourn times.

1. INTRODUCTION

The inversion of the Laplace transform for the (unconditional) sojourn time V in the M/M/1 queue with egalitarian processor sharing (PS) was carried out by Morrison [1] and also by Yashkov [2, pp.75–78]. The inversion had been obtained by means of deforming the contour of integration in the complex plane to one around a cut between two branch points. It led to an integral representation for $\mathsf{P}(\mathsf{V} > \mathsf{x})$. Later similar procedure was carried out by Flatto [3] for the M/M/1 queue subject to the random order of service (RO) discipline. This result was used by Flatto to find the exact asymptotics for the tail of the stationary distribution waiting time in the M/M/1—RO queue. Recently the same result from [3] has been used by Lebedev [4] to obtain the stochastic estimates of the normalized maxima of the waiting times in the M/M/1—RO queue. In this note, we carry the results of analysis of [3, 4] a step further, and show that the known results of [3, 4] for the M/M/1—RO queue carry over to the M/M/1—PS queue, and vice versa. In particular, it concerns the asymptotic tail behaviour of the normalized maxima of the waiting times and the sojourn times in the RO and the PS queues, respectively.

The remainder of the article is organized as follows. Section 2 contains necessary preliminaries. The relevant theorems are reflected in Section 3. Two variants of the proof of main result (Theorem 3.5) are also given in this section. Finally, Section 4 contains our short conclusion.

2. PRELIMINARIES

Jobs arrive at a single server queue according to a Poisson process with rate $\lambda > 0$. The required service times are i.i.d. random variables with the distribution $B(x) = 1 - \mathrm{e}^{-\mu \mathrm{x}}$ which has the mean $1/\mu < \infty$. We assume that $\rho = \lambda/\mu < 1$, this assumption insures stability. Under the PS discipline, the service rate is equally shared among all jobs present. In other words, when there are $n \geq 1$ jobs present, each job receives service at rate $1/n$ (see [2] or [5] for additional details). Under the RO discipline, the jobs are served in random order: whenever the server becomes free, the next job

[1] This research was partially supported by Grant no. Sci.Sch.–934.2003.1 (Head R.A.Minlos) and Grant to the Program of Fundamental Researches of Russian Academy of Sciences (the Division of Informatics) "New physical and structural solutions in infotelecommunications"(Head N.A.Kuznetsov).

to be served is selected uniformly at random from the jobs present, if any. Thus, each job being equally likely to be chosen.

Let V be the (unconditional) stationary sojourn time for a particular job in the PS queue, and W be the stationary waiting time in the RO queue. To distinguish the disciplines, we shall sometimes use the marks PS or RO.

3. THE RESULTS

The following theorem has been known already since mid eighties (see, for example, Formula (2.20) of [1] or Formula (2.49) of [2] or Formula (4.4) of [5]).

Theorem 3.1.

$$\mathsf{P}(V_{PS} > x) = \frac{2}{1-\rho}\int_0^{\pi} \frac{\mathrm{e}^{-\gamma\left[2\rho^{1/2}-(1+\rho)\cos\gamma\right]/[(1-\rho)\sin\gamma]-\mu x(1-\rho)^2/\left(1+\rho-2\rho^{1/2}\cos\gamma\right)}}{1+\mathrm{e}^{-\pi\left[2\rho^{1/2}-(1+\rho)\cos\gamma\right]/[(1-\rho)\sin\gamma]}}\sin\gamma \mathrm{d}\gamma. \tag{3.1}$$

Remark 31. Theorem 3.1 had been proved two decades ago in [1, 2] for the case when time was normalized such that $\mu = 1$. Here it is re-formulated for the generic case.

Theorems 3.2 and 3.3 had been proved by Flatto [3] one decade ago.

Theorem 3.2.

$$\mathsf{P}(\mathsf{W}_{\mathsf{RO}} > \mathsf{x}) = \frac{2(1-\rho)}{\rho}\int_0^{\pi} \frac{\mathrm{e}^{[2\phi(\theta)-\theta]\cot\theta}}{\mathrm{e}^{\pi\cot\theta}+1}\,\frac{\mathrm{e}^{-\left[1-2\rho^{-1/2}\cos\theta+\rho^{-1}\right]\lambda\mathsf{x}}}{\left(1-2\rho^{-1/2}\cos\theta+\rho^{-1}\right)^2}\sin\theta \mathrm{d}\theta, \tag{3.2}$$

where

$$\phi(\theta) = \arctan\frac{\sin\theta}{\cos\theta-\rho^{1/2}}, \quad 0 \le \phi(\theta) \le \pi. \tag{3.3}$$

Remark 32. In [3], time is normalized such that arrivals occur at unit rate, but here it is re-formulated for the generic case.

It is difficult to obtain insight into $\mathsf{P}(\mathsf{V}_{\mathsf{PS}} > \mathsf{x})$ and $\mathsf{P}(\mathsf{W}_{\mathsf{RO}} > \mathsf{x})$ directly from Theorems 3.1 and 3.2. Nevertheless, the satisfactory breakthroughs in this direction have occurred in [1] and [3], respectively. We point out the following result.

Theorem 3.3.

$$\mathsf{P}(\mathsf{W}_{\mathsf{RO}} > \mathsf{x}) \sim \frac{\mathsf{c}_1\mathrm{e}^{-\left[\mathrm{c}_2\lambda\mathrm{x}+\mathrm{c}_3(\lambda\mathrm{x})^{1/3}\right]}}{\left(\lambda\mathsf{x}\right)^{5/6}}, \quad \mathsf{x}\to\infty, \tag{3.4}$$

where

$$c_1 = 2^{2/3}3^{-1/2}\pi^{5/6}\rho^{17/12}\frac{1+\rho^{1/2}}{\left(1-\rho^{1/2}\right)^3}\mathrm{e}^{\left(1+\rho^{1/2}\right)/\left(1-\rho^{1/2}\right)},$$

$$c_2 = \left(\rho^{-1/2}-1\right)^2,$$

$$c_3 = 3\left(\frac{\pi}{2}\right)^{2/3}\rho^{-1/6}.$$

Theorem 3.3 has been used in [4] to prove the following claims.

Let W_n be the waiting time of n–th arriving job, $n = 1, 2, \ldots$ under the RO discipline. We consider $M_n = \max\{W_1, W_2, \ldots, W_n\}$ as $n \to \infty$. It is clearly that the process $\{W_n\}$ is regenerative with respect to the sequence of the points when the consecutive busy periods are initiated. Now we assume that $\lambda = 1$. We shall denote the right–hand side of (3.4) as $W^\circ(x)$.

Let b_n be the numerical consequence of normalizing constants such that $nW^\circ(b_n) \to 1$ as $n \to \infty$. Then, taking into account some simple facts from extreme value theory, it follows from Theorem 3.3 (in the situation when $\lambda = 1$, in this case $1/\mu = \rho \in (0, 1)$):

Theorem 3.4.

$$(i) \qquad \liminf_{n\to\infty} \mathsf{P}\left(\mathsf{c_2}\left(\mathsf{M_n} - \mathsf{b_n}\right) \le \mathsf{x}\right) \ge \mathrm{e}^{-\mathrm{e}^{-\mathsf{x}}},$$

where

$$b_n = \frac{1}{c_2}\left[\ln n - c_3\left(\frac{\ln n}{c_2}\right)^{1/3} + \frac{5}{6}\ln\left(\frac{\ln n}{c_2}\right) + \ln c_1\right], \quad n \ge 2,$$

and constants c_1, c_2 and c_3 are given in Theorem 3.3;

$$(ii) \qquad \limsup_{n\to\infty} \frac{M_n}{\ln n} \le \frac{1}{c_2}$$

.

Theorems 3.2 — 3.4 are valid only for the M/M/1—RO queue. Now we show how these theorems are carried over to the M/M/1—Ps queue. It holds

Theorem 3.5. *Taking into account the equivalence relation*

$$\rho\mathsf{P}(\mathsf{V_{PS}} > \mathsf{x}) = \mathsf{P}(\mathsf{W_{RO}} > \mathsf{x}) = \rho\mathsf{P}(\mathsf{W_{RO}} > \mathsf{x}|\mathsf{W_{RO}} > 0), \tag{3.5}$$

the Theorems 3.2 — 3.4 directly yield the exact and asymptotic tail behaviour of the sojourn time distribution in the M/M/1—PS queue.

Proof. The first way. The proof relies only on the fact that the sojourn time in the GI/M/1 queue under the PS discipline is equal in distribution to the waiting time under the RO discipline for a job arriving to a non–empty RO system. This important observation was done by Cohen [6] who, starting from Ramaswami's results [7], showed that the transform of the delay distribution satisfies in both cases the same differential equation (with an unique solution). See for details Yashkov [2, Theorem 2.8]. It leads to (3.5). In other words, the sojourn time V_{PS} is closely related to the conditional waiting time $W_{RO}|W_{RO} > 0$ in the GI/M/1 queue through a simple multiplicative constant, namely,

$$\mathsf{P}(\mathsf{V_{PS}} > \mathsf{x}) = \frac{1}{\xi}\mathsf{P}(\mathsf{W_{RO}} > \mathsf{x}), \quad \mathsf{x} \ge 0,$$

where $\xi = \mathsf{P}(\mathsf{W_{RO}} > 0)$. We note that ξ coincides with the probability $\mathsf{P}(\mathsf{L} > 0)$, where L is the number of jobs in the queue GI/M/1 at arrival epochs, and ξ reduces to ρ for the case of a Poisson input. □

Proof. The second way. The equivalence relation (3.5) may be derived from Theorems 3.1 and 3.2. More exactly, (3.5) follows after re–formulation of Theorem 3.1 to the form of Theorem 3.2 in a spirit of transformations which resemble ones from [2, pages 75–77]. To this end, it is useful to use some elementary facts from complex analysis and trigonometry. We shall not go into details. □

Remark 33. Deeper connection between processor sharing and random order of service has been observed already in [8, Remark 2].

4. CONCLUSION

Theorem 3.5 demonstrates how Theorems 3.2 — 3.4 are carried over to the M/M/1—PS queueing system. This leads to some new results concerning the exact and asymptotic tail behaviour of the sojourn time distribution in the M/M/1 processor sharing queue.

REFERENCES

1. Morrison J. Response time for a processor–sharing system. *SIAM J. Appl. Math.*, 1985, vol. 45, no. 1, pp. 152–167.
2. Yashkov S.F. *Analysis of Queues in Computers*. Moscow: Radio i Svyaz, 1989 (in Russian).
3. Flatto L. The waiting time distribution for the random order service M/M/1 queue. *Ann. Appl. Prob.*, 1997, vol. 7, no. 2, pp.382–409.
4. Lebedev A.V.Waiting time maxima in an M/M/1 random order service queue. *Problems of Information Transmission*, 2005, vol. 41, no. 3, pp.123–127 (in Russian).
5. Yashkov S.F. Processor–sharing queues: some progress in analysis (invited paper). *Queueing Systems*, 1987, vol. 2, no. 1, pp.1–17.
6. Cohen J.W. On processor sharing and random service (letter to the editor). *J. Appl. Prob.*, 1984, vol. 21, no. 4, pp.937–937.
7. Ramaswami V. The sojourn time in the GI/M/1 queue with processor sharing. *J. Appl. Prob.*, 1984, vol. 21, no. 2, pp.437–442.
8. Yashkov S.F. A derivation of response time distribution for an M/G/1 processor–sharing queue. *Problems of Control and Information Theory*, 1983, vol. 12, no. 2, pp. 133–148.

This paper was recommended for publication by V.I.Venets, a member of the Editorial Board