Информационные процессы, Том 4, № 3, 2004, стр. 251–255.

A NOTE ON A HEAVY TRAFFIC LIMIT THEOREM FOR THE M/D/1—FBPS QUEUE[1]

S.F.Yashkov*, **A.S.Yashkova****

Institute for Information Transmission Problems, 19, Bolshoi Karetny lane, 101447 Moscow GSP–4, Russia. E-mail: yashkov@iitp.ru

**Municipal Institute of Zhukovsky, Chair Appl. Informatics in Economics, 15, Mayakovsky Street, 140180 Zhukovsky, Moscow region, Russia.*

Received May 31, 2004

Abstract—In this short article we give a simple proof of the heavy traffic limit theorem for the sojourn time distribution, suitably scaled, in the M/D/1 queueing system under the foreground–background processor sharing (FBPS) discipline. The FBPS discipline gives the preemptive–resume priority in service to those jobs that have received the least amount of service so far. This theorem was announced without the proof in [5, Th. 6.7].

1. INTRODUCTION

We consider the M/D/1 queue with the foreground–background processor sharing (FBPS) discipline. This discipline gives the preemptive–resume priority in service to those jobs that have received the least amount of service time. If there are $1 \leq n < \infty$ such jobs, then they are served simultaneously at rate $1/n$. The problem of the determination of the steady–state sojourn time distribution for the M/G/1—FBPS queue has been solved independently by Kleinrock [1] and Yashkov [2], [3] (see [4, Appendix, Th. 1]). They assumed that the distribution function of the service requirements (better: the lengths of jobs) $B(x)$ is absolutely continuous. Because of a weak continuity argument, their results hold for the case when $B(x)$ has no density, in particular, for the M/D/1—FBPS. The purpose of this short article is to find the limiting distribution of the steady–state sojourn time in this system, suitably scaled, when the offered load tends to one. We present a simple analytical proof of such heavy–traffic limit theorem.

The remainder of this article is organized as follows. Section 2 contains necessary preliminaries. The proof of main result is given in Section 3. Finally, in Section 3 we state our conclusion.

2. PRELIMINARIES

The FBPS discipline can be described with some details in the following manner. A job with attained service (the age) a does not receive service unless there are no jobs in the system with the age less than a. The single processor simultaneously serves those and only those with the least age, but each at a rate $\frac{1}{n}$ if the number of the youngest jobs is $n,\ 1 \leq n < \infty$. Thus a job (or a set of jobs) with the least amount of attained service (e.g., a new arrival) has the highest preemptive–resume priority which decreases in accordance with an increment of its age. Jumps in the service rate occur at the epochs of arrivals, departures and when the age of the jobs which share processor reaches to the age of some interrupted jobs. See [1], [4] and [5] for additional information on the FBPS discipline.

We consider the M/D/1 queue where the jobs arrive according to a Poisson process with rate λ. The service discipline is the FBPS described above. The job's lengths, i.e., the service requirements, are i.i.d.

[1] The research of S.F.Y. was partially supported by Grant no. Sci.Sch.–934.2003 (Head R.A.Minlos).

random variables with the deterministic distribution (the job's lengths are constant)

$$B(x) = \begin{cases} 0, & 0 \le x < u, \\ 1, & x \ge u \end{cases}$$

that has the Laplace–Stieltjes transform (LST) $\beta(s) = \exp(-su)$ and the moments $\beta_i = u^i, \quad i = 1, 2, \ldots$. The offered load is equal to $\rho = \lambda u < 1$. We will use V to denote the steady–state sojourn time of a job in the queue M/D/1—FBPS.

The starting point will be the Theorem 1 from [4, Appendix] describing (in more general set-up) the LST $v(s) \doteq \mathsf{E}[\mathrm{e}^{-sV}]$ of the sojourn time distribution, which gives us the following decomposition of V

$$V \stackrel{d}{=} G + \Pi. \tag{2.1}$$

The equality (2.1) is rewritten in terms of the LST as $v(s) = \gamma(s)\pi(s)$, where

$$\gamma(s) \doteq \mathsf{E}[\mathrm{e}^{-sG}] = w(s + \lambda - \lambda\pi(s)) \tag{2.2}$$

and $\pi(s) \doteq \mathsf{E}[\mathrm{e}^{-s\Pi}]$ is the unique positive solution of the well–known functional equation for the LST of the busy period distribution [6]

$$\pi(s) = \beta(s + \lambda - \lambda\pi(s)) \tag{2.3}$$

with the smallest absolute value. The LST $w(s)$ in (2.2) is given by the celebrated Pollaczek–Khintchine formula

$$w(s) = \frac{s(1-\rho)}{s + \lambda - \lambda\beta(s)}.$$

It follows easily from the expressions above that $\mathsf{E}[V] = \mathsf{E}[G] + \mathsf{E}[\Pi]$, where

$$\mathsf{E}[G] = \gamma_1 = \mathsf{E}[W]/(1-\rho) = \lambda\beta_2/[2(1-\rho)^2], \tag{2.4}$$

$$\mathsf{E}[W] = \lambda\beta_2/[2(1-\rho)], \tag{2.5}$$

$$\mathsf{E}[\Pi] = \pi_1 = u/(1-\rho). \tag{2.6}$$

3. MAIN RESULT

We will prove main theorem following the line sketched in [7].

Theorem 3.1. *For the M/D/1—FBPS queue,* $\lim_{\rho\uparrow 1} \mathsf{P}(V/\mathsf{E}[V] \le x)$ *has the density*

$$q(x) = (\pi x)^{-1/2} \exp(-x/4) - [\mathrm{erfc}(x^{1/2}/2)]/2, \tag{3.1}$$

where

$$\mathrm{erfc}(y) = 2\pi^{-1/2} \int_y^\infty \exp(-t^2)\, dt. \tag{3.2}$$

Proof. The basic step in the proof is to show that, when $\rho \uparrow 1$, the random variable $G/\mathsf{E}[G]$ in (2.1) converges in distribution to the random variable Q with the density $q(x)$, that is,

$$G/\mathsf{E}[G] \stackrel{d}{\to} Q \qquad \text{for } \rho \uparrow 1, \tag{3.3}$$

where $\mathsf{E}[G]$ is given by (2.4). To this end, we rewrite (2.3) to the form

$$\pi(s) = \beta(s + \frac{s\rho}{1-\rho} f_\Pi(s)) \tag{3.4}$$

taking into account (2.6). Here and further

$$f_\Pi(s) = \frac{1-\pi(s)}{s\pi_1}. \tag{3.5}$$

This is the LST of the residual busy period distribution. Another probability term for it is the LST of the excess of Π (that is, the LST of the random modification of Π). By analogy with (3.5), it holds the following expression for the distribution of the residual length of a job

$$f_B(s) = \frac{1-\beta(s)}{s\beta_1}. \tag{3.6}$$

The first moments of the excess for the random variables Π and B will be denote as $f_{1\Pi}$ and f_{1B}, respectively. For example, it is well known that

$$f_{1B} = \beta_2/(2\beta_1). \tag{3.7}$$

Taking into account (2.3), one can rewrite (2.2) to the form

$$\gamma(s) = 1 - \rho + \rho f_\Pi(s), \tag{3.8}$$

where $f_\Pi(s)$ is given by (3.5). Now the equation (3.4) can be represented in terms of $\gamma(s)$ as follows

$$\pi(s) = \beta\left(\frac{s\gamma(s)}{1-\rho}\right).$$

Taking into account (3.5), (3.6) and the last equality, the equation (3.8) can be reduced to the form

$$\gamma(s) = \frac{1-\rho}{1-\rho f_B\left(\frac{s\gamma(s)}{1-\rho}\right)}. \tag{3.9}$$

Let $\varepsilon = 1-\rho, \quad \varepsilon \ll 1$. If we rewrite the equality (3.9) in terms of $\varphi_\varepsilon(s) \doteq \gamma(\varepsilon^2 s)$, then we get

$$\varphi_\varepsilon(s) = [1 + (1-\varepsilon) f_{1B} s \varphi_\varepsilon(s) \alpha_\varepsilon(s)]^{-1}. \tag{3.10}$$

Here f_{1B} is given by (3.7) and

$$\alpha_\varepsilon(s) \doteq \mathsf{E}[\mathrm{e}^{-sA}] = \frac{1 - f_B(\varepsilon s \varphi_\varepsilon(s))}{f_{1B}\varepsilon s \varphi_\varepsilon(s)}.$$

Since the function $\alpha_\varepsilon(s)$ is analitical in the half–plane $\mathrm{Re}\, s > 0$, continuous up to $\mathrm{Re}\, s = 0$ and $|\alpha_\varepsilon(s)| \le 1$, then the value of $\alpha_\varepsilon(s)$ at $s = it$ coincides with the value of the characteristic function for the same random variable A, namely, $\mathsf{E}[\mathrm{e}^{itA}]$ at the point $-t$ for all $\varepsilon > 0$. Hence $\varphi_\varepsilon(s)$ satisfies the quadratic equation

$$(1-\varepsilon) f_{1B} s \alpha_\varepsilon(s) \varphi_\varepsilon^2(s) + \varphi_\varepsilon(s) - 1 = 0.$$

The solution of this equation is given by

$$\varphi_\varepsilon^(s) = \frac{2}{1 + \sqrt{1 + 4(1-\varepsilon) f_{1B} s \alpha_\varepsilon(s)}}. \tag{3.11}$$

Because of $\lim_{\varepsilon\downarrow 0}\varphi_\varepsilon(s) = 1$, we have to choose the positive sign in the denominator of (3.11).

Let $\varepsilon \downarrow 0$, then $\alpha_\varepsilon(s) \to 1$ for all $0 \le s < \infty$. In virtue of this, the equality (3.11) yields

$$\lim_{\varepsilon\downarrow 0}\gamma(\varepsilon^2 s) = \frac{2}{1+\sqrt{1+4f_{1B}s}}, \qquad 0 \le s < \infty.$$

This means that

$$\lim_{\rho\uparrow 1}\gamma\left(\frac{(1-\rho)^2 s}{f_{1B}}\right) = \frac{2}{1+\sqrt{1+4s}}, \qquad \mathrm{Re}\, s \ge 0. \tag{3.12}$$

Now it is not difficult to invert the Laplace transform that stands in the right–hand side of (3.12). It yields the right–hand side of (3.1) together with (3.2). Thus we proved the equality (3.3). In other words,

$$\frac{2(1-\rho)^2\beta_1}{\beta_2}G \xrightarrow{d} Q \qquad \text{for } \rho \uparrow 1, \tag{3.13}$$

where the density of the distribution of the random variable Q is given by (3.1).

It remains to note that $(1-\rho)^2\Pi \xrightarrow{d} 0$ as $\rho \uparrow 1$. Then assertion of Theorem 3.1 follows from (3.13) and (2.1). □

Theorem 3.1 is interesting because, for the limiting distribution of the sojourn time scaled by its expectation, a distribution other than exponential appears. This is the important difference in comparison with the classical limit theorems for the FCFS queue (see, for example, Cohen [8] and Whitt [9]) and also with the limit theorem for the conditional sojourn time distribution in the M/G/1 queue under egalitarian processor sharing (see [7], [10] and [11]).

We can see from the right–hand side of (3.12) that our LST is very closely related to the LST of the (scaled) canonical distribution of the reflected Brownian motion (RBM) starting off empty (see, for example, Section 7 in Abate, Choudhury, Lucantoni and Whitt [12], in particular, (7.2) of the cited paper).

Corollary 3.1. *The moments of the random variable Q are*

$$q_n = (2n)!/(n+1)!, \quad n = 1, 2, \ldots$$

Proof. For simplicity, it is sufficient to use (7.3) from [12]. □

Remark 3.1. For the case deterministic distribution of $B(x)$ considered above, the discipline FBPS coincides with the discipline LRPT (Longest Remaining Processing Time). The LRPT policy assumes that, at every moment of time, only the set of jobs with the maximal remaining service time share the server in pure (egalitarian) processor sharing fashion. Thus, the LRPT discipline biases towards the longest jobs.

Remark 3.2. Other interesting treatments of the M/G/1 queue under various service disciplines can be found in Kleinrock [1, Ch. 4], Yashkov [5], [13], Cohen [8, Pt. 4], Cooper [14] and Conway et al. [15, Ch. 8].

It is well known that $\mathsf{E}[V]$ in the M/G/1—FCFS queue is minimized for the case M/D/1—FCFS if one imposes the constrains that λ and β_1 are fixed. Taking into account Theorems 2 and 3 from [4, Appendix], now it is easily to show that, in sharp contrast to the FCFS, $\mathsf{E}[V]$ in the M/G/1—FBPS queue is maximized for the case M/D/1—FBPS under the same constrains. Theorem 3.1 provides the tight upper bound for the family of the all normalized versions of the sojourn time distributions in the work–conserving M/G/1 queue (in particular, for any processor sharing discipline). The upper bound is well suitable for numerical calculations.

4. CONCLUSION

We gave the simple analytical proof of the heavy traffic limit theorem for the distribution of the sojourn time, scaled by its expectation, in the M/D/1 queue under foreground–background processor sharing discipline. This theorem determines the tight upper bound for the set of all normalized versions of the sojourn time distributions in the work–conserving M/G/1 queue as the offered load increases to one. The upper bound have of practical interest and it is well suitable for numerical studies.

REFERENCES

1. Kleinrock L. *Queueing Systems*. New-York: Wiley, 1976, vol. 2. Russian edition: Kleinrock L. *Computer Systems with Queues.* Ed. Tsybakov B.S. Moscow: Mir, 1979.
2. Yashkov S.F. On foreground–background sharing a processor among jobs with minimal serviced length. *Tekhnika Sredstv Svyazi. Ser. ASU*, 1978, no. 2, pp. 51–62 (in Russian).
3. Yashkov S.F. Analysis of a queueing system with multilevel priority discipline of time–sharing. In: *Cybernetics Problems in Automatic Control Systems of Technology Processes*. Moscow: MDNTP, 1978, pp. 138–144 (in Russian).
4. Yashkov S.F., Yashkova A.S. The time–dependent solution of the M/G/1—FBPS queue. *Information Processes*, 2004, vol. 4, no. 2, 175–187 (available at http://www.jip.ru/2004/175–187.pdf).
5. Yashkov S.F. Mathematical problems in the theory of shared-processor systems. In: *Itogi Nauki i Tekhniki. Ser.: Probability Theory*. Moscow: VINITI, 1990, vol. 29, pp. 3–82 (in Russian). Engl. edition: *J. of Soviet Mathematics*, 1992, vol. 58, no. 2 (Jan.), pp. 101–147.
6. Takacs L. *Introduction to the Theory of Queues.* New York: Oxford Univ. Press, 1962.
7. Yashkov S.F. Some limit theorems for processor sharing queueing systems. In: *Long–Range Tools of Telecommunication and Integrated Communication Systems.* Ed. Kuznetsov N.A. Moscow: IITP, 1992, pt. 1. pp. 214–220 (in Russian).
8. Cohen J.W. *The Single Server Queue.* (2nd ed.) Amsterdam: North–Holland, 1982.
9. Whitt W. Heavy traffic limit theorems for queues: a survey. In: *Mathematical Methods in Queueing Theory. Ser.: Lect. Notes in Econ. and Math. Systems.* Ed. Clarke A.B. Berlin: Springer, 1974, vol. 98. pp. 307–350.
10. Yashkov S.F. On a heavy traffic limit theorem for the M/G/1 processor–sharing queue. *Commun. in Statist. — Stochastic Models*, 1993, vol. 9, no. 3, pp. 467–471.
11. Sengupta B. An approximation for the sojourn–time distribution for the GI/G/1 processor–sharing queue. *Commun. in Statist. — Stochastic Models*, 1992, vol. 8, no. 1, pp. 35–58.
12. Abate J., Choudhury G.L., Lucantoni D.M., Whitt W. Asymptotic analysis of tail probabilities based on the computation of moments. *The Annals of Appl. Probability*, 1995, vol. 5, no. 4, pp. 983–1007.
13. Yashkov S.F. *Analysis of Queues in Computers*. Moscow: Radio i Svyaz, 1989 (in Russian).
14. Cooper R.B. *Introduction to Queueing Theory.* (2nd ed.) New York: North–Holland, 1981.
15. Conway R.W., Maxwell V.L., Miller L.W. *Theory of Scheduling.* Reading MA: Addison–Vesley, 1967. Russian edition: Conway R.W., Maxwell V.L., Miller L.W. *Theory of Scheduling.* Ed. Basharin G.P. Moscow: Nauka, 1975.

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