Информационные процессы, Том 4, № 3, 2004, стр. 269–274.

MATHEMATICAL MODELS

SOME EXTENSION OF THE HEAVY TRAFFIC LIMIT THEOREM FOR THE M/G/1—EPS 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 September 13, 2004

Abstract—We extend the heavy traffic limit theorem for the stationary sojourn time distribution, suitably scaled, in the M/G/1 egalitarian processor sharing (EPS) queue from [12], [13] to the case when the M/G/1—EPS queue is modified by having $K \geq 0$ extra permanent jobs with infinite sizes.

1. INTRODUCTION

Egalitarian Processor Sharing (EPS) queueing systems first became popular by the works of Kleinrock [1] and Yashkov [2], and were originally proposed by Kleinrock in 1967 to analyze the performance of time–sharing schedulung algorithms in computers. Nowadays, the processor sharing paradigm has emerged as a powerful concept for modeling of Web servers. In particular, the EPS has also become relevant in modeling elastic traffic, the flow–level performance of bandwidth–sharing protocols in the nodes of modern computer–communication networks, etc. (The Transmission Control Protocol (TCP) in adaptive window mechanism can be considered as an example.)

In the EPS discipline, the processor (server) is shared equally by all jobs in the system. To put more concretely, when $1 \leq n < \infty$ jobs are present in the system, each job receives service at rate $1/n$. In other words, all these jobs receive $1/n$ times the rate of service which a solitary job in the processor would receive. Jumps of the service rate occur at the instants of arrivals and departures from the system. Therefore, the rate of service received by a specific job fluctuates with time and, importantly, its sojourn time depends not only on the jobs in the processor at its time of arrival there, but also on subsequent arrivals shorter of which can overtake a specific job. This makes the EPS system intrinsically much harder to analyze than, say, the M/G/1 queue with the First Come — First Served (FCFS) or other classical disciplines.

The problem of the exact determination of the stationary sojourn time distribution in the M/G/1—EPS queue was open a long time. This problem was first solved, after puzzling researchers for 15 years, by Yashkov in [3],[2] in terms of double Laplace tranforms (LT). Another approach to the exact solution of the same problem was proposed by Schassberger in [4]. (Later some additional contributions to this problem were also made by Sengupta, van den Berg [5], Grishechkin, Whitt [6] and Nunez–Queija [7] among others). From a probabilistic point of view, processor sharing queues are very interesting in view of their connections with (non-trivial) branching processes (like the processes by Crump–Mode–Jagers). Nevertheless, the deep investigation of the EPS queues is based on specific analytic methods introduced, in essence, in [3], [2]. We refer also to [8] for the deep study and the state of the art in this area. Concerning recent breakthroughs to the transient (time–dependent) analysis of the M/G/1-EPS queue see [9], [10] (and also [8], [11]) and the references therein. Almost all available at present analytic solutions of the M/G/1—EPS system (and also

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

many new) can be derived as special cases of our transient results which are obtained in [9] for more general situation when the M/G/1—EPS queue is modified by having $K \geq 0$ extra permanent jobs with infinite sizes.

The purpose of this article is to extend the heavy traffic limit theorem for the stationary sojourn time distribution, suitably scaled, in the M/G/1—EPS queue from [12], [13] to this case when there are $K \geq 0$ extra permanent jobs. The case of $K = 0$ was also studied independently by Sengupta [14] at the same time as it was done in [12], and later also by Zwart in 2000.

The remainder of the article is organized as follows. Section 2 contains necessary preliminaries. The proof of main result is given in Section 3. We note as important consequence that the scaled unconditional sojourn time distribution has the form of the Mellin–Stieltjes convolution of two independent distributions: the $(K+1)$–phase Erlangian distribution and the service time distribution $B(x)$. Finally, Section 4 contains our conclusion.

2. PRELIMINARIES

We consider the stationary egalitarian processor sharing queue M/G/1 with the intensity λ of Poisson input process (of the standard jobs) and service time (standard job's size) distribution $B(x) \quad (B(0+) = 0,\ B(\infty) = 1)$ with the mean $\beta_1 < \infty$ and the Laplace–Stieltjes transform (LST) $\beta(r)$. When there are $n \geq 1$ jobs in the EPS queueing system then each of them receives service at a rate which is $1/n$ times the rate of service that a solitary job in the system would receive. The offered load is equal to $\rho = \lambda\beta_1 < 1$.

The M/G/1—EPS queue is modified by having $K \geq 0$ extra permanent jobs with infinite sizes. Note that the behaviour of this model is independent of the service time distribution of the permanent jobs because of such jobs are always in service. We denote the conditional sojourn time of the standard (non permanent) job with the size u as $V_K(u)$ and put $v_K(r,u) \doteq \mathsf{E}[\mathrm{e}^{-rV_K(u)}]$ (u is vieved as a parameter). As before, ρ is the offered load to the system per unit of time due to the standard (the Poisson) jobs.

Our starting point will be the Proposition 4.1 from [9] describing the LST $v_K(r,u)$ of the sojourn time distribution. Here we rewrite it as

Theorem 2.1. *For $\rho < 1$ and $K \geq 0$, it holds*

$$v_K(r,u) \doteq \mathsf{E}[\mathrm{e}^{-rV_K(u)}] = v(r,u)^{K+1}, \tag{2.1}$$

where $v(r,u)$, the LST of the conditional sojourn time of the standard job (with the size u) in the M/G/1—EPS queue without permanent jobs (that is, when $K = 0$, then there are only the standard jobs), is given by

$$v(r,u) \doteq \mathsf{E}[\mathrm{e}^{-rV(u)}] = \frac{(1-\rho)\delta(r,u)}{1 - \tilde{a}(r,0,u)/\psi(r,u)}\,. \tag{2.2}$$

Here

$$\tilde{a}(r,0,u) = \lambda \int_0^u \psi(r,u-x)\mathrm{e}^{-x(r+\lambda)}(1-B(x))dx + \lambda \mathrm{e}^{-u(r+\lambda)} \int_u^\infty (1-B(x))dx, \tag{2.3}$$

$$\delta(r,u) = \frac{\mathrm{e}^{-u(r+\lambda)}}{\psi(r,u)} \tag{2.4}$$

and $\psi(r,u)$ is the LST w.r.t. x (argument r) of some unknown function $\Psi(x,u)$ of the two variables, which, in turn, is given by its LT w.r.t. u (argument q)

$$\tilde{\psi}(r,q) = \frac{q + r + \lambda\beta(q+r+\lambda)}{(q+r+\lambda)(q+\lambda\beta(q+r+\lambda))}, \quad r \geq 0,\ q > -\lambda\pi(r). \tag{2.5}$$

We note that $\pi(r)$ in $q > -\lambda\pi(r)$ (in (2.5)) is the LST of the busy period distribution, i.e., it is the positive root of the celebrated Takacs functional equation

$$\pi(r) = \beta(r + \lambda - \lambda\pi(r)) \tag{2.6}$$

with the smallest absolute value.

Remark 2.1. The formula for $v(r, u)$ in Theorem 2.1 above is equivalent to the Theorem 4 from [2] (except for a difference in notations).

3. THE RESULT

Taking into account (2.3) and (2.4), we can rewrite (2.2) as

$$v(r, u) = \frac{(1-\rho)\delta(r, u)}{1 - \lambda\int_0^\infty \varphi(r, x, u)(1 - B(x))\,dx}, \tag{3.1}$$

where

$$\varphi(r, x, u) = \begin{cases} \delta(r, u) & \text{for } x \geq u, \\ \delta(r, u)/\delta(s, u - x) & \text{for } x < u. \end{cases} \tag{3.2}$$

Remark 3.1. The following equivalent form of (3.2) can sometimes be used to describe some important extensions of Theorem 2.1 (although they are not discussed here)

$$\varphi(r, x, u) = \mathrm{e}^{-(x\wedge u)(r+\lambda)+\lambda\int_0^{x\wedge u}\varphi_B(r,u-y)\,dy}, \quad x \in [0, \infty),$$

where

$$\varphi_B(r, t) \doteq \int_0^\infty \varphi(r, x, t)\,dB(x) = \int_0^t \mathrm{e}^{-\int_{t-x}^t (r+\lambda-\lambda\varphi_B(r,y))\,dy}\,dB(x) + (1 - B(t))\mathrm{e}^{-\int_0^t (r+\lambda-\lambda\varphi_B(r,y))\,dy}.$$

Here the last expression is the functional equation for the LST of the terminating (at time t) busy period in the M/G/1—EPS queue [3], [8]. The solution of this equation has been obtained in terms of the function $\psi(r, t) \doteq \exp(-\lambda\int_0^t \varphi_B(r, y)\,dy)$ (more precisely, in terms of the LT w.r.t. t of $\psi(r, t)$ — see (2.5)).

Further we shall use the notation

$$\delta_j(u) = \lim_{r\downarrow 0}(-1)^j \frac{\partial^j \delta(r, u)}{\partial r^j},$$

$$\varphi_j(x, u) = \lim_{r\downarrow 0}(-1)^j \frac{\partial^j \varphi(r, x, u)}{\partial r^j}, \quad j = 1, 2, \ldots.$$

Let $\varepsilon = 1 - \rho$, $\varepsilon << 1$. Replacing r in (2.1) (and hence, in (2.2)) by εr and using the Tailor series expansion in a point εr for small $\varepsilon > 0$, it follows that

$$v_K(r, u) = \left(\frac{\varepsilon\left[1 - \varepsilon r\delta_1(u) + \frac{\varepsilon^2 r^2}{2!}\delta_2(u) - \ldots\right]}{1 - (1-\varepsilon)\left[1 - \varepsilon r\overline{\varphi}_1(u) + \frac{\varepsilon^2 r^2}{2!}\overline{\varphi}_2(u) - \frac{\varepsilon^3 r^3}{3!}\overline{\varphi}_3(u) + \ldots\right]}\right)^{K+1}. \tag{3.3}$$

Here we used the notation

$$\overline{\varphi}_j(u) = \beta_1^{-1}\int_0^\infty \varphi_j(x, u)(1 - B(x))\,dx, \quad j = 1, 2, \ldots \tag{3.4}$$

where $\varphi_j(x,u)$ is given above.

Taking into account that

$$\mathsf{E}[V_K(u)] = \frac{(K+1)u}{1-\rho} = (K+1)\delta_1(u) + \frac{\rho\overline{\varphi}_1(u)}{1-\rho}, \quad K = 0,1,2,\ldots$$

(see, for example, the formula (2.66) from [8], which, however, is stated only for the case $K = 0$, but can easily be extended to $K > 0$, as it is done in [9, Corollary 4.1]), the similar chain of the inferences as in [12], [13] can be used to prove the following theorem.

Theorem 3.1. *For the M/G/1—EPS queue with $K \geq 0$ permanent jobs, $\beta_1 < \infty$ and any fixed $u \in [0,\infty)$,*

$$\lim_{\rho\uparrow 1} \mathsf{P}(V_K(u)(1-\rho)/u \leq x) = \left(1 - \mathrm{e}^{-x}\right)^{(K+1)*}, \; x \geq 0, \tag{3.5}$$

where $(K+1)*$ is the symbol of the $(K+1)$–fold Stieltjes convolution of the corresponding distribution function with itself.

Proof. See arguments above and the proof of the main theorem in [13] (we omit the details). □

Thus, the limiting distribution in the right–hand side of (3.5) is the $(K+1)$–phase Erlangian distribution with the mean $(K+1)$. Now the main theorem from [13] follows as the special case for $K = 0$.

Remark 3.2. Theorem 3.1 can be easily re–formulated in other equivalent forms.

It is clearly that the LST of the unconditional (on u) sojourn time distribution is given by

$$v_K(r) \doteq \mathsf{E}[\mathrm{e}^{-rV_K}] = \int_0^\infty v_K(r,u)\, dB(u). \tag{3.6}$$

Now we have the following consequence

Corollary 3.1.

$$\lim_{\rho\uparrow 1} v_K(r(1-\rho)) = \int_0^\infty \left(\frac{1}{1+ru}\right)^{K+1} dB(u). \tag{3.7}$$

Proof. The result follows directly from (3.6), Theorem 3.1 above and bounded convergence theorem. □

Remark 3.3. The formula (3.7) is the so–called Mellin–Stieltjes convolution (see [15]) of two independent distributions: the $(K+1)$–phase Erlangian distribution and $B(x)$. In other words, the limiting distribution is the product of two corresponding independent random variables.

Remark 3.4. We note that the Mellin–Stieltjes convolution (notation is $\overset{S}{*}$) of some distribution function $A(x)$ and $B(x)$ $\quad(x \in \mathbf{R}^+)$ is coincides with the usual Stieltjes convolution (notation is $*$) of the distribution functions $A(\mathrm{e}^x)$ and $B(\mathrm{e}^x)$. In other words,

$$C(x) \doteq A \overset{S}{*} B(x) = \int_0^\infty A(x/y)dB(y) = A * B(\mathrm{e}^x).$$

These results allow us to use the well known properties of the product of two random variables (see, for example, Feller [16]) with the purpose to find explicit expressions for the limiting distributions of the scaled sojourn time. It has of practical interest, in particular, in the case of heavy-tailed distributions $B(x)$ (more precisely, in the case of subexponential $B(x)$, for example, with regularly varying tails (at infinity)). In particular, an old result of Breiman (1965) that is hidden in [15, Theorem 8.15.3], says that if the random variable B has the distribution function $1 - B(x) \sim x^{\alpha}\ell(x),\ \alpha > 0,\ x > 0$, where $\ell(x)$ is slowly varying function at infinity, and A is another random variable independent of B satisfying $\mathsf{E}[A^{\gamma}] < \infty$ for some $\gamma > \alpha$, then

$$\mathsf{P}(AB > x) \sim \mathsf{E}[A^{\alpha}]\mathsf{P}(B > x), \quad x \to \infty. \tag{3.8}$$

Due to (3.8), the expressions above are well suitable for numerical calculations in the heavy traffic case. For example, if $B(x)$ has a Pareto distribution, then the heavy–traffic limiting distribution of the sojourn time belongs (for the case $K = 0$) to the class of Pareto Mixtures of Exponentials (PME), introduced in Abate et al. [17]. We will not give the details here since ones can easily be retrieved using, say, the results of [17] or by Shiryaev [18, Ch. 4] as supplement to the corresponding assertions above.

Remark 3.5. The results of the time–dependent exact solutions on the sojourn time and queue–length distributions in the M/G/1 queue under three main disciplines (EPS, FBPS and LCFS with preemptions) allow us also to obtain a number of other limit theorems — not only in the heavy traffic but even in overloaded mode. Here we do not discuss these problems in details. Note only that the time–dependent queue–length distribution in the M/G/1—EPS queue (under zero initial condition, $K = 0$ and in terms of the triple transforms) coincides with that for the M/G/1 queue with preemptive LCFS discipline. Such solution for the M/G/1—EPS queue is known at least with 1988 (see, for example, the formula (2.108) in [8, p. 100] or the Theorem 3.2 [11] that holds for any $0 < \rho < \infty$ in spite of the restriction $\rho < 1$ in the corresponding statement [11, p. 202]). Thus, all transient performance measures for the number of jobs at time t in the M/G/1 queue must coincide for the EPS and LCFS with preemptions. Of course, it is not true for the time-dependent (and stationary, too) sojourn time distributions for these disciplines. We note also that the time–dependent queue-length distribution in the M/G/1—FBPS queue has the form which sharply differs from the EPS [8, Ch. 2, 3]. Concerning the optimal properties of the FBPS discipline see, for example, Avrachenkov, Ayesta et al. [19] that supplements earlier results of Yashkov (1978), reflected, in particular, in [8].

4. CONCLUSION

We gave the simple analytical proof of the heavy traffic limit theorem for the distribution of the sojourn time, suitably scaled, in the M/G/1 queue under egalitarian processor sharing discipline with $K \geq 0$ permanent jobs. This theorem represents some extension of the results [12], [13] and also it leads to interesting new consequences which are convenient for obtaining explicit expressions well suitable for numerical calculations. We pointed out also some trends of recent investigations.

REFERENCES

1. Kleinrock L. *Queueing Systems*. New-York: Wiley, 1976, vol. 2. Russian edition: Kleinrock L. *Computer systems with Queues*. Moscow: Mir, 1979.

2. 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 (Publ. House of the Hungarian Acad. of Sci., Budapest).

3. Yashkov S.F. Some results of analyzing a probabilistic model of remote processing systems. *Avtomatika i Vychislitel'naya Tekhnika.*, 1981, no. 4, pp. 3-11 (in Russian). Engl. edition by Allerton Press (New York, USA): *Automatic Control and Computer Sciences*, 1981, vol. 15, no. 4, pp. 1-8.

4. Schassberger R. A new approach to the M/G/1 processor-sharing queue. *Advances in Applied Probability*, 1984, vol. 16, no. 1, pp. 202-213.

5. van den Berg J.L. *Sojourn Times in Feedback and Processor–Sharing Queues.* PhD thesis. Utrecht Univ., 1990.

6. Whitt W. The M/G/1 processor–sharing queue with long and short jobs. *Unpublished manuscript*, 1998 (Sept.).

7. Nunez–Queija R. *Processor Sharing Models for Integrated Services Networks.* PhD thesis. Eindhoven Univ. of Technology, 2000.

8. Yashkov S.F. *Analysis of Queues in Computers*. Moscow: Radio i Svyaz, 1989 (in Russian).

9. Yashkova A.S., Yashkov S.F. Distribution of the virtual sojourn time in the M/G/1 processor sharing queue. *Information Processes*, 2003, vol. 3, no. 2, pp. 128–137 (available at http://www.jip.ru/2003/128–137.pdf).

10. Yashkov S.F. On sojourn time problem in processor sharing queue. In: *Int. Conf. "Kolmogorov and Contemporary Mathematics". Abstracts.* (Moscow, June 16–21, 2003). Moscow: Moscow State Univ., 2003, pp. 594–595.

11. Yashkov S.F. Time–dependent analysis of processsor–sharing queue. In: *Queueing, Performance and Control in ATM. Proc. 13th Int. Teletraffic Congress.* (Copenhagen, June 19–26, 1991). Eds. J.W.Cohen and C.D.Pack. Amsterdam: Elsevier, 1991, pp. 199-204.

12. 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).

13. Yashkov S.F. On a heavy traffic limit theorem for the M/G/1 processor–sharing queue. *Communications in Statistics — Stochastic Models*, 1993, vol. 9, no. 3, pp. 467–471.

14. Sengupta B. An approximation for the sojourn–time distribution for the GI/G/1 processor–sharing queue. *Communications in Statistics — Stochastic Models*, 1992, vol. 8, no. 1, pp. 35–58.

15. Bingham N.H., Goldie C.M., Teugels J.L. *Regular Variation.* (Encycl. of Math. and its Appl., vol. 27). Cambridge: Cambridge Univ. Press, 1987.

16. Feller W. *Introduction to Probability Theory and its Applications.* New York: Wiley, 1966, vol 2. Russian Edition: Feller W. *Introduction to Probability Theory and its Applications.* Ed. Prokhorov Yu.V. Moscow: Mir, 1967, vol. 2.

17. Abate J., Choudhury G.L., Whitt W. Waiting–time tail probabilities in queues with long–tail service –time distributions. *Queueing Systems*, 1994, vol. 16, pp. 311–338.

18. Shiryaev A.N. *Essentials of Stochastic Finance Mathematics.* Moscow: Phasis, 1998, vol. 1: Facts. Models, vol. 2: Theory (in Russian). Engl. edition: Shiryaev A.N. *Essentials of Stochastic Finance: Facts, Models, Theory.* Singapore: World Scientific, 1999.

19. Avrachenkov K.E., Ayesta U., Brown P., Nyberg E. Differentiation between short and long TCP flows: predictability of the response time. In: *Proc. of IEEE Infocom'2004*. Hong Kong, 2004.

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