Информационные процессы, Том 6, № 3, 2006, стр. 237–249.

MATHEMATICAL MODELS

THE MOMENTS OF THE SOJOURN TIME IN THE M/G/1 PROCESSOR SHARING SYSTEM [1]

S.F.Yashkov

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

Received June 15, 2006

Abstract—We give a short review of some recent achievements in analysis of the M/G/1 queue with egalitarian processor sharing discipline (EPS). The paper contains readly applicable expressions for the j-th moments ($j \in \mathbb{N}$) of the (conditional) stationary sojourn time in the M/G/1—EPS queue with K ($K \in 0 \cup \mathbb{N}$) permanent jobs of infinite size. We show also how to simplify the computations of the moments.

1. INTRODUCTION

Processor sharing queues, made very attractive models by the works of Kleinrock [1], [2] and Yashkov [3], [4], play a central role in queueing theory. These models were originally proposed to analyze the performance of scheduling algorithms in time–sharing computer systems, and continue to find new applications which pose interesting mathematical problems. Over the past few years, the processor sharing paradigm has emerged as a powerful concept for modeling of Web servers, in particular, for evaluating the flow–level performance of end–to–end flow control mechanisms like Transmission Control Protocol (TCP) in Internet.

The mathematical analysis of processor sharing queues has resulted in many insightful results. Yet, a number of challenging problems remains to be explored. The main goal of this paper is to gain understanding the problem of the moments of the stationary sojourn time in the M/G/1 queue with egalitarian processor sharing (EPS), and to derive the formulas for the j-th moments ($j \in \mathbb{N}$) of the (conditional) sojourn time in the M/G/1—EPS queue with K ($K \in 0 \cup \mathbb{N}$) permanent jobs of infinite size. Our results complement and develop the corresponding sections of the paper by Yashkova and Yashkov [5].

An idea of the EPS discipline[2] was introduced by Kleinrock [1] who studied only M/M/1 case as a limit of the round–robin queue. In particular, he first showed that the mean sojourn time conditioned on the initial job size (service requirement of the job) is linear function of the size of the job. For an overview of the literature on processor–sharing queueing systems we refer to Kleinrock [2] (1976), Koboyashi and Konheim [6] (1977), Jaiswal [7] (1982), Yashkov [4] (1987), [8] (1990) and Yashkova and Yashkov [5] (2003).

[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).

[2] Under 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 harder to analyze than, say, the First Come — First Served (FCFS) queue.

The exact determination of the stationary sojourn time distribution in the M/G/1—EPS queue was an open problem for a long time. After puzzling researchers for 15 years, Yashkov [9] (1981), [3] (1983) found an analytic solution of this problem in terms of double Laplace tranforms (LT) (all details contains also his book [10] (1989)). Schassberger [11] (1984) provided another (completely different) approach to the exact solution by considering the EPS discipline as a limit of the round–robin model (in discrete time). Later similar solutions were also made by means of the variants of the methods from [3] and from [11] (or their combinations). See, for example, Rege and Sengupta [12] (1994), Whitt [13] (1998) or Ward and Whitt [14] (2000). In particular, Ward and Whitt give some additional explanation of our results and apply them to predicting congestions in the M/G/1—EPS queue. Here we do not have possibility to discuss the contributions of other authors (for example, Brandt and Brandt [15] (2003), Asare and Foster[16] (1983), Grishechkin, et alii) to the closely related problems. We only mention that the EPS queue with permanent jobs has been also studied in [13, 17] from point of view which is different from our approach. A telecommunication system with CPU scheduling under SCO–UNIX can be considered as an example of using the EPS model with permanent jobs for its description and predicting delays of the jobs [17].

In fact, our method has turned out to be a very fruitful to derive many further results, for example, the time-dependent queue-length and sojourn time distributions in this and related models (see, for example, [10, 8, 18, 5][3]).

These results hold for any stability condition. Besides, the entire transient and equilibrium behaviour of the M/G/1—EPS queue is contained in the results mentioned, and the most (if not all) available at present analytic solutions (and also many new) can be derived from them as special cases via standard arguments (for example, by means of the Abel's/Tauber's theorems). However, we shall not consider the transient solutions in this paper.

The rest of the paper is organized as follows. In Section 2 we introduce some notations and describe our starting point represented by Theorem 2.1. In Section 3 we obtain some interesting consequences of Theorem 2.1, some of which were proved earlier as self–contained theorems but now ones are derived as special cases. The final section contains few closing remarks.

2. PRELIMINARIES

In this section we give a short review of the M/G/1—EPS queue with K permanent jobs (only in steady state). For the time-dependent results we refer to [5] and also to [10, 8].

Jobs arrive to the single processor (server) according to a Poisson process with the rate $\lambda > 0$. Their sizes (reguired service times) are i.i.d. random variables with a general distribution function $B(x)$ ($(B(0) = 0,\ B(\infty) = 1)$) with the mean $\beta_1 < \infty$ and the Laplace–Stieltjes transform (LST) $\beta(s)$. Let β_j denote the j–th moment of $B(x)$, $j \in \mathbb{N}$. The service discipline is the EPS: every job

[3] Indeed, the assumptions which required to use the steady–state (stationary) solutions of any queueing systems are rarely satisfied in real life. To be able really to apply queueing results in design and analysis of technical systems, in very many cases, the obtained results of steady state analysis are not sufficient. For example, it is often necessary to investigate the behaviour of the queue while it progresses towards a steady state (if and when a steady state exists). Even the average queue length at time t gives us much more information in comparison with the stationary mean of the number of jobs.

However, few stochastic systems are known to have exact time–dependent (transient) solutions for the distributions of the processes. As a rule, such systems are the M/G/1 queues with simpler disciplines (for example, FCFS, see, for example, Takacs [19]). Besides, all time–dependent solutions of the queues of the type M/G/1 are obtained in terms of double transforms (on space and time) from which it is very hard to extract necessary information concerning the behaviour of the system. (Moreover, much more advanced mathematical techniques become necessary for the time–dependent solutions in comparison with steady state analysis.) Some exceptions give variations of the M/M/1—FCFS queue for which closed–form transient solutions are known. As a rule, the exact transient analysis of the M/M/1—FCFS queue involves infinite sums of Bessel functions. In general, explicit exact solutions are highly unlikely for the time–dependent cases.

is being served with rate $1/n$, when $n > 0$ jobs are present in the system. The EPS discipline is modified by having $K \geq 0$ extra permanent jobs with infinite sizes. The system works in steady state. In other words, $\rho = \lambda\beta_1 < 1$ and very long time went from the instant 0 that marks the start of the work of our system till current time.

It is well known, due to Sakata et al. [20], that the stationary distribution $(\mathsf{P}_n)_{n\geq 0}$ of the number of ordinary jobs in the M/G/1—EPS queue as $K = 0$ is geometrically distributed

$$\mathsf{P}_{0\mathsf{n}} = (1-\rho)\rho^{\mathsf{n}}, \quad \mathsf{n} \in \mathsf{0} \cup \mathbb{N}, \tag{2.1}$$

where $\rho = \lambda \int_0^\infty (1 - B(x))dx < 1$. We note that $(\mathsf{P}_{0\mathsf{n}})_{\mathsf{n}\geq 0}$ depends on the service time only through its mean.

For $K \geq 0$ the equality (2.1) takes the form

$$\mathsf{P}_{\mathsf{K}\mathsf{n}} = (1-\rho)^{\mathsf{K}+1}\binom{\mathsf{n}+\mathsf{K}}{\mathsf{K}}\rho^{\mathsf{n}}, \quad \mathsf{n} \in \mathsf{0} \cup \mathbb{N}. \tag{2.2}$$

We shall let that $V_K(u)$ denotes the conditional sojourn time of a job of the size u upon its arrival. This job enter into the EPS system with $K \geq 0$ permanent jobs in steady state. Let $v_{Kj} = \mathbb{E}[V_K(u)^j]$. (We shall omit the index K in these and similar notations when $K = 0$.)

Define the LST of $V_K(u)$ by $v_K(r,u) = \mathbb{E}[\mathrm{e}^{-rV_K(u)}]$ for $\mathrm{Re}\, r \geq 0$ and $u \geq 0$.

Let $\pi(r)$ be the LST of the busy period distribution (due to ordinary, that is, non–permanent jobs). In other words, it is the positive root of the well–known Takács functional equation [19]

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

with the smallest absolutely value.

It is known from [5] the following theorem

Theorem 2.1. *When $\rho < 1$,*

$$v_K(r,u) \doteq \mathbb{E}[\mathrm{e}^{-rV_K(u)}] = v(r,u)^{K+1}, \tag{2.4}$$

where $v(r,u)$ is given by the the equality (2.5):

$$v(r,u) \doteq \mathbb{E}[\mathrm{e}^{-rV(u)}] = \frac{(1-\rho)\mathrm{e}^{-u(r+\lambda)}}{\psi(r,u) - \tilde{a}(r,0,u)}\,. \tag{2.5}$$

Here

$$\tilde{a}(r,0,u) = \lambda\psi(r,u) * \left[\mathrm{e}^{-u(r+\lambda)}(1-B(u))\right] + \lambda\mathrm{e}^{-u(r+\lambda)}\int_u^\infty (1-B(x))dx, \tag{2.6}$$

where "$*$"is the Stieltjes convolution sign (on variable u), and $\psi(r,u)$ is the LST (with respect to x) of some function $\Psi(x,u)$ of two variables (possessing the probability density on variable x), which, in turn, has a Laplace transform (LT) with respect to u(argument q)

$$\tilde{\psi}(r,q) = \frac{q + r + \lambda\beta(q + r + \lambda)}{(q + r + \lambda)(q + \lambda\beta(q + r + \lambda))} \quad (r \geq 0,\, q > -\lambda\pi(r)). \tag{2.7}$$

In (2.7), $\beta(r) = \int_0^\infty \mathrm{e}^{-rx}\, dB(x)$ and $\pi(r)$ (in the conditions imposed on (2.7)) is understood as the minimal solution of the functional equation (2.3).

Thus, the function $\tilde{\psi}(r,q)$ is given in the form of the two–dimensional transform of the function $\Psi(x,u)$)

$$\tilde{\psi}(r,q)=\int_0^\infty\int_0^\infty \mathrm{e}^{-rx-qu}d_x\Psi(x,u)du. \tag{2.8}$$

In other words, $\psi(r,u)$ in equality (2.5) is the Laplace transform inversion operator, $\psi(r,u)=\mathcal{L}^{-1}(\tilde{\psi}(r,q))(r,u)$, that is, the contour Bromvich integral

$$\psi(r,u)=\frac{1}{2\pi i}\int_{-i\infty+0}^{+i\infty+0}\tilde{\psi}(r,q)\mathrm{e}^{qu}\,dq.$$

Remark 21. Briefly, we have derived the expression for $\mathbb{E}[\mathrm{e}^{-rV_K(u)}]$ by writing the sojourn time as some generalized functional on a branching process (like the processes by Crump–Mode–Jagers) by means of simple extensions of (non–trivial) arguments from [9, 3]. Using the structure of the branching process, we found and solved a system of partial differential equations (of the first order) determining the components of a (non–trivial, too) decomposition of $V_K(u)$. It leads to $\mathbb{E}[\mathrm{e}^{-rV_K(u)}]$ (see also Remark 33).

3. RESULTS

We showed in the Section 2 that the determination of the steady–state sojourn time distribution in the queue M/G/1—EPS with K permanent jobs is simple extension of the results from [3], [9]. However, the solution contains the Bromwich countour integrals. First we consider the case $K=0$. Equivalent form of (2.5) (without contour integrals) is given in the following theorem.

Theorem 3.1. *Equivalent form of (2.5) (without the Bromwich countour integrals) is given by*

$$\frac{1}{v(r,u)}=\sum_{n=0}^\infty\frac{r^n}{n!}\xi_n(u), \tag{3.1}$$

where

$$\xi_0(u)=1,\quad \xi_n(u)=\frac{n}{(1-\rho)^n}u^{n-1}*W^{(n-1)*}(u),\quad n=1,2,\dots \tag{3.2}$$

Here $W^{(n-1)*}(u)$ is $(n-1)$–fold convolution of the steady–state waiting time distribution $W(u)$ in the familiar M/G/1—FCFS system with itself ($W^{0*}(u)=\mathbf{1}(u)$, $W^{1*}(u)=W(u)$), the LST of $W(u)$ is given by the well–known Pollaczek–Khintchine formula as

$$w(q)=\frac{1-\rho}{1-\rho f(q)}, \tag{3.3}$$

where $f(q)=(1-\beta(q))/(q\beta_1)$ is the LST of the excess of $B(\cdot)$, that is, $F(x)=\beta_1^{-1}\int_0^x(1-B(y))dy$ ($F^{0*}(x)=\mathbf{1}(x)$, the Heaviside function, $F^{1*}(x)=F(x)$).

Proof. We rewrite (2.5) in the form of Theorem 3.2 from [11] (see also (5.5) in [4]), namely

$$v(r,u)=\frac{(1-\rho)\delta(r,u)}{1-\rho\delta(r,u)\left[\int_0^u\frac{dF(x)}{\delta(r,u-x)}+(1-F(u))\right]}\quad(\mathrm{Re}\ r\ge 0), \tag{3.4}$$

where

$$\delta(r,u)=\mathrm{e}^{-u(r+\lambda)}/\psi(r,u) \tag{3.5}$$

and $F(x)$ is introduced in Theorem 3.1. To reach our aim, it is used the LT of $1/\delta(r,u)$ with respect to u (argument q), which is found from (2.7) as $\tilde{\psi}(r,q-r-\lambda),\quad r\ge 0,\ q>r+\lambda-\lambda\pi(r)$ (cf.

also the third line on p.8 in [4]). Now we obtain after simple algebra the following power series expansion of the LT of the function $1/v(r,u), \quad r \geq 0,\ u \geq 0$

$$\int_0^\infty \mathrm{e}^{-qu}\frac{1}{v(r,u)}du = \frac{1}{q}\left[1+\frac{1}{1-\rho}\frac{r}{q}\frac{1}{1-\frac{1}{1-\rho}\frac{r}{q}w(q)}\right]$$
$$= \frac{1}{q}\left[1+\sum_{n=1}^{\infty}\left(\frac{1}{1-\rho}\frac{r}{q}\right)^n w(q)^{n-1}\right], \tag{3.6}$$

where $w(q)$ is given by (3.3). We note that $\left|\frac{rw(q)}{(1-\rho)q}\right| < 1$ as $q > r+\lambda-\lambda\pi(r),\ \rho < 1$. Now it is easily to invert analytically (on argument q) each term of the power series in r (3.6). The result is given by (3.2) whence it follows (3.1), the right–hand side of which is the power series in r with coefficients $\xi_n(u)/n!$. □

The idea of such approach goes back to Heaviside. Similar results are obtained in [21], [22]. In fact, the form of $\mathsf{Var}[V(u)]$ [9], [3] (see the equality (3.10) below) stimulates a guess about the possibility of such expansion.

Remark 31. The formula for $W^{n*}(x)$ in (3.2) can be represented in the following form

$$W^{n*}(x) = (1-\rho)^n \sum_{k=0}^{\infty}\binom{k+n-1}{n-1}\rho^k F^{k*}(x).$$

It is done, for example, by inversion of $w(q)^n$, where $w(q)$ is given by (3.3).

Remark 32. We note that the by–product of our analysis is the distribution function $W(x)$ whose LST is given by (3.3). However, the analysis of EPS queue gives the other quantity (corresponding to a non–probability measure) $W^\circ(x) = W(x)/(1-\rho)$. The form of the LST of $W^\circ(x)$ is well–known: $w^\circ(q) = \sum_{n=0}^{\infty}\rho^n f^n(q)$. Unlike $W(x)$, $W^\circ(x)$ is well defined for all $\rho > 0$ and $x > 0$. It can be shown that $W^\circ(x) < \infty$ for all $\rho > 0$, $x > 0$ and for any $B(\cdot)$ (despite on the fact that, for $\rho \geq 1$, $W^\circ(x) \to \infty$ as $x \to \infty$).

Theorem 3.2. *Let $v_n(u) = \mathbb{E}[V(u)^n]$, $n = 1, 2, \ldots$ Then it holds the following recursive formula*

$$v_n(u) = \sum_{i=1}^{n}\binom{n}{i}v_{n-i}(u)\xi_i(u)(-1)^{i+1} \tag{3.7}$$

Proof. Because $v(r,u)$ is analytical function in r (in particular, in $r = 0$), we can use the Tailor series expansion of $v(r,u)$ for small $r > 0$

$$v(r,u) = 1 - \frac{r}{1!}v_1(u) + \frac{r^2}{2!}v_2(u) - \frac{r^3}{3!}v_3(u) + \ldots \tag{3.8}$$

The product of (3.8) and (3.1) gives

$$-\frac{r}{1!}[v_1(u)-\xi_1(u)] + \frac{r^2}{2!}[v_2(u) - 2v_1(u)\xi_1(u) + \xi_2(u)]$$
$$-\frac{r^3}{3!}[v_3(u) - 3v_2(u)\xi_1(u) + 3v_1(u)\xi_2(u) - \xi_3(u)] + \ldots = 0$$

and it leads to (3.7) after differentiating n times with respect to r and setting $r = 0$. □

In particular, the expressions for the first two moments of $V(u)$ are:

$$v_1(u) = \mathbb{E}[V(u)] = u/(1-\rho) \tag{3.9}$$

(this is well-known result due to Sakata et al. [20] (1969)),

$$\mathsf{Var}[V(u)] = v_2(u) - v_1^2(u) = \frac{2}{(1-\rho)^2}\int_0^u (u-x)(1-W(x))\,dx, \tag{3.10}$$

where $W(x)$ is introduced in Theorem 3.1, and it is expressed as

$$W(x) = (1-\rho)\sum_{n=0}^{\infty}\rho^n F^{n*}(x) \tag{3.11}$$

(other variables were introduced above).

The formula (3.2) implies that $\xi_1(u) = \mathsf{E}[V(u)]$ in (3.9). The formula for the conditional variance (3.10) was first obtained by Yashkov [9]. The standard way for the computation of the moments is the following

$$v_n(u) = \lim_{r\downarrow 0}(-1^n)\frac{\partial^n v(r,u)}{\partial r^n}, \quad n \in \mathbb{N}. \tag{3.12}$$

However, the LST $v(r,u)$ in Theorem 2.1 is very hard to differentiate in r more than once (practically almost impossible matter) since this LST has a rather complex form due to a highly complicated form of such constituents of (2.5) as $\tilde{a}$ and ψ. Therefore $\mathsf{Var}[V(u)]$ is first obtained by solving an alternative system of differential equations (see, for example, [10, Chapter 2] and also Appendix) which are derived by analogy with the equations of [9, Section 2] or with the equations in the proof of [3, Theorem 4]. These equations are simpler forms of equations from [9], [3] because ones are composed not for the LST $v(r,u)$ but only for the second and the first moments. Thus the formula (3.7) for the case $n = 2$ in Theorem 3.2 was derived 20 years earlier than the same formula for arbitrary integer n (see [21] (1999), [22] (2000))

It can be useful for asymptotic expansion of $v_n(u)$ for small and large u in the spirit of such expansion for $\mathsf{Var}[V(u)]$ (see final section for some details). Such results for $\mathsf{Var}[V(u)]$ were obtained at first in [9] (1981) (see also [3]).

Now we show how to extend the formulas (3.9) and (3.10) to the case when the M/G/1—EPS queue is modified by having $K \geq 0$ extra permanent jobs with infinite sizes.

Theorem 3.3. *In the setting above,*

$$\mathbb{E}[V_K(u)] = \frac{(K+1)u}{1-\rho}, \tag{3.13}$$

$$\mathsf{Var}[V_K(u)] = \frac{2(K+1)}{(1-\rho)^2}\int_0^u (u-x)(1-W(x))dx, \tag{3.14}$$

where $W(u)$ is the steady–state waiting time distribution in the M/G/1—FCFS queue, represented by the equality (3.11).

Proof. In our case, the equality (3.12) takes the form

$$v_{Kn}(u) = \lim_{r\downarrow 0}(-1^n)\frac{\partial^n v_K(r,u)}{\partial r^n}, \quad n \in \mathbb{N}. \tag{3.15}$$

The formula (3.13) follows directly from (2.4) by means of applying (3.15) as $n = 1$.

Taking into account (3.10), the formula (3.14) follows also from (2.4) by means of applying (3.15) as $n = 2$ after some simple algebra. □

Remark 33. An alternative way to obtain (3.14) is the following. We can compose and solve the system of the partial differential equations (of the first order) which satisfy the second and the first moments of $V_K(u)$. The variant of such equations is known from [8, 10] as $K = 0$. We point out the following fact. These equations rely on a decomposition of the sojourn time of the (tagged)job with the size u that arrives to the EPS queue when n standard jobs are present with remaining service demands $x_1, \dots x_n$ (a key ingredient of analysis). Denoting this conditional sojourn time by $V_{Kn}(u; x_1, \dots x_n)$, it holds

$$V_{Kn}(u; x_1, \dots x_n) \stackrel{d}{=} (K+1)D(u) + \sum_{i=1}^{n} \Phi(x_i, u), \tag{3.16}$$

where all components are independent random variables.

The random variable $D(u)$ constitutes a "main" component of the sojourn time: it has the distribution of the sojourn time of a job with the size u that enters into a empty (from the standard jobs) system. By the way, its LST is given by (3.5). When the system is not empty, the i–th standard job (among the jobs which are sharing the capacity of the processor together with permanent jobs), having remaining size x_i, "adds" a delay $\Phi(x_i, u) = \Phi(x_i \wedge u, u)$ to the new job's sojourn time. Note that $D(u) = \Phi(x_i, u)$ for $x_i \geq u$. Then the same chain of arguments as in [3] can be used to derive (2.4).

4. CONCLUSION

Using Theorems 2.1 and 3.2, we can easily obtain all other moments of $V_K(u)$ in M/G/1—EPS queue with $K \geq 0$ permanent jobs. However, the exact expression even for the variance of the sojourn time (see (3.14) involves an integration term, making an exact computations difficult from practical point of view. The same holds for the third, fourth, etc. moments. This difficulty remains also in the case $K = 0$. To overcome the difficulty, it is possibly to obtain some simple approximations for the second moments, see, for example, Villela et al. [23]. We note that there exist also an upper (and lower) bounds for $\mathsf{Var}[V(u)]$. These bounds only depends on ρ and the size of job u. In addition, the bounds have the attractive property of intensitivity to $B(x)$, and the difference between the upper and lower bounds is small, particularly, for small and moderate values of ρ. These second moments tight bounds can be easily generalized into higher moments of $V(u)$ and also to the case $K > 0$.

It are also known the asymptotic estimates of $\mathsf{Var}[V(u)]$ as $u \to 0$ and $u \to \infty$ [24]. For example,

$$Var[V(u)] \sim \frac{u^2\rho}{(1-\rho)^2} \quad \text{as} \quad u \to 0.$$

This is some asymptotics of the sojourn time variance of a very small jobs, and it leads to intensitive upper bounds with special structure requiring only knowledge of the traffic load and the job size. Now our results may be easily extended to the higher moments and also to the case K permanent jobs. Moreover, some preliminary analysis of asymptotics (see some examples in [9, 10]) tells us about a high accuracy of such estimates in many typical cases.

APPENDIX

The appendix contains the derivation of the first two moments of the random variable $V(u)$ for $K = 0$ (see [10] for details).

Expressions for the first two moments.

The components of (3.16) are the analogs of the "delays elements" of the tagged job in the case of the M/M/1 queue with a round–robin discipline [25], which are interpretable in terms of terminating busy periods. Since the case in question is considerably more complex in virtue of the arbitrary distribution $B(x)$ and the simultaneous servicing with varying rate, to calculate the distribution functions of the components of (3.16), it is necessary to solve some system of differential equations. This question will be considered a bit later.

Let us define $\delta_j(u) = \mathbb{E}[D(u)^j]$ and $\varphi_j(x,u) = \mathbb{E}[\Phi(x,u)^j]$ for $j \in \mathbb{N}$. It then follows from (3.16) that as $K = 0$

$$\mathbb{E}[V(u)|n; x_1, \dots x_n] = \delta_1(u) + \sum_{i=1}^{n} \varphi_1(x_i, u), \tag{A.1}$$

$$\begin{aligned}\mathbb{E}[V(u)^2|n; x_1, \dots x_n] = \delta_2(u) + \sum_{i=1}^{n} \varphi^2(x_i, u) \\ + 2\delta_1(u) \sum_{i=1}^{n} \varphi_1(x_i, u) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \varphi_1(x_i, u)\varphi_1(x_j, u).\end{aligned} \tag{A.2}$$

We know that the probability of dwelling in the system $L = n$ jobs whose remaining sizes lie in infinitesimally small neighborhoods of the points $x_1, \dots x_n$ has the form of (2.1), where $\rho = \lambda \int_0^\infty (1 - B(x))dx < 1$.

We obtain the unconditional moments of the random variable $V(u)$ by averaging (A.1) and (A.2) over the distribution (2.1)

$$v_1(u) = \delta_1(u) + \frac{\rho}{1-\rho}\bar{\varphi}_1(u), \tag{A.3}$$

$$v_2(u) = \delta_2(u) + \frac{\rho}{1-\rho}\bar{\varphi}_2(u) + \frac{2\rho}{1-\rho}\delta_1(u)\varphi_1(u) + \frac{2\rho^2}{(1-\rho)^2}\bar{\varphi}_1^2(u). \tag{A.4}$$

Here, we used the notation

$$\bar{\varphi}_i(u) = \frac{1}{\beta_1} \int_0^\infty \varphi_i(x,u)(1 - B(x))\mathrm{d}x. \tag{A.5}$$

Computation of $v_1(u)$ and $v_2(u)$.

Let us prove (3.16) as $K = 0$. Without loss of generality, we may assume that a tagged job of size u arrives in the system at the initial instant and that the system is in the state $n; x_1, \dots x_n$, $n \in \mathbb{N}$. The random variable $V(u)$ is the sum of the serviced parts of the sizes of those jobs that were serviced the same time as the tagged job. Since these sizes are independent, to prove (3.16), we need only to show that it is possible to group them in a suitable way. Branching processes generated by the n jobs present in the system at the instant of appearance of the tagged job and by the marked job itself were examined for this purpose in [25, 3, 9]. Processes of such type arise in queueing theory in connection with the derivation of the functional equation for the busy period distribution [19].

In describing branching processes, it is useful to go over to a new time scale: intervals of time throughout which n jobs are serviced in the system are compressed to $1/n$ their original sizes. With such a random change of time scale, the rate of servicing is no longer a random variable but is constant and equal to unity. However, the input flow then changes. The memoryless property remains, but the intensity becomes variable, depending on the number of jobs being serviced. In other words, on an interval of time for which there are n jobs in the system, the intensity of appearance is equal to λn (and to λ for $n = 0$); the input flow on this interval can be regarded as the superposition of independent Poisson flows of intensity λ.

We now introduce branching processes of generation of descendants of jobs. To each job we assign an independent Poisson process of intensity λ. This flow is "included" at the instant of appearance of a job in the system and it is "excluded" at the instant of departure of a job from the system. We shall consider jobs of this flow to be descendants of the original job. All arriving jobs can be partitioned according to a transitive descendance relationship into disjoint classes according to to the number of jobs in the system at the initial instant. Since the instants of appearance and the sizes of jobs of different classes are independent, the question of computation of the distribution $V(u)$ can be reduced to calculating the distributions of certain functionals of the independent branching processes introduced.

Consider an interval J of the length u located arbitrarily on the time axis. Suppose that, at the initial instant of that interval, there is the job of size x in the system. We define the random variable $\Phi(x,u)$ as the sum of these parts of sizes of that job and its descendants that are serviced in the interval J. We should point out that, in the case of an infinite interval ($u \to \infty$), the random variable $\Phi(x,u)$ becomes the usual busy period for which the size of job opening it is fixed and equal to x.

It follows from the definition that $\Phi(x,u)$ is, for $x \geq u$, independent of x. In this case, it is convenient to introduce for it the special notation $D(u) = \Phi(x,u)$ for $x \geq u$. For the moments, we obviously have

$$\delta_j(u) = \varphi_j(x,u) \qquad \text{for } x \geq u, \quad j \in \mathbb{N}. \tag{A.6}$$

The exposition given above enables us to obtain (3.16) directly (for $K = 0$).

For the moments of the random variables $D(u)$ and $\Phi(x,u)$, we can derive the differential equations, examining the intervals of time differing by an infinitesimal amount Δ and remembering that each newly appearing job can be regarded as an ancestor generating a branching process of the type indicated, that is independent of the other jobs.

Specifically,

$$D(u+\Delta) = \begin{cases} \Delta + D(u), & \text{if no descendants appears,} \\ \Delta + D(u) + \Phi(x,u), & \text{if a descendant of size } x \text{ appears.} \end{cases} \tag{A.7}$$

Similarly, for $x < u$,

$$\Psi(x+\Delta, u+\Delta) = \begin{cases} \Delta + \Phi(x,u), & \text{if no descendants appears,} \\ \Delta + \Phi(x,u) + \Phi(x,u), & \text{if a descendant of size } x \text{ appears.} \end{cases} \tag{A.8}$$

If we apply to (A.7) and (A.8) the formula for the total mathematical expectation, we obtain, by passing to the limit, the following two systems of equations, represented below together with the obvious additional conditions

$$\frac{d\delta_1(u)}{du} = 1 + \lambda \int_0^\infty \varphi_1(y,u)\,dB(y), \tag{A.9}$$

$$\frac{\partial \varphi_1(x,u)}{\partial x} + \frac{\partial \varphi_1(x,u)}{\partial u} = 1 + \lambda \int_0^\infty \varphi_1(y,u)\,dB(y), \tag{A.10}$$

$$\delta_1(0) = \varphi_1(0,u) = \varphi_1(u,0) = 0$$

.

$$\frac{d\delta_2(u)}{du} = 2\delta_1(u)\left[1 + \lambda \int_0^\infty \varphi_1(y,u)\,dB(y)\right] + \lambda \int_0^\infty \varphi_2(y,u)\,dB(y), \tag{A.11}$$

$$\frac{\partial\varphi_2(x,u)}{\partial x}+\frac{\partial\varphi_2(x,u)}{\partial u}=2\varphi_1(x,u)\left[1+\lambda\int_0^\infty\varphi_1(y,u)dB(y)\right]+\lambda\int_0^\infty\varphi_2(y,u)dB(y), \qquad (A.12)$$

$$\delta_2(0)=\varphi_2(0,u)=\varphi_2(u,0)=0$$

.

The system (A.9), (A.10) is easily solved in the domain $x<u$. To do this, it is necessary to solve the following partial differential equation of the first order

$$\frac{\partial\varphi_1}{\partial x}+\frac{\partial\varphi_1}{\partial u}=\frac{\partial\delta_1}{\partial u}.$$

For this equation, the conjugate system of ordinary differential equations have the form $dx/1+du/1=d\varphi_1/\delta_1'$. The first integrals are $C_1=u-x$, $C_2=\varphi_1-\delta_1$, and the general solution is expressed as $C_2=f(C_1)$. From here, taking into account the additional conditions, it follows

$$\varphi_1(x,u)=\delta_1(u)-\delta_1(u-x) \qquad \text{for } x<u. \qquad (A.13)$$

Substituting (A.13) into (A.9) and taking into account the equality (A.6), we obtain the following equation

$$\frac{d\delta_1(u)}{du}=1+\lambda\delta_1(u)-\lambda\int_0^u\delta_1(u-y)dB(y).$$

Applying the Laplace transform to each term of this equation yields

$$\int_0^\infty \mathrm{e}^{-ru}\delta_1(u)du=\frac{1}{r^2[1-\lambda(1-\beta(r))/r]}. \qquad (A.14)$$

It follows from (A.14)

$$\delta_1(u)=u+\int_0^\infty(u-x)\sum_{n=1}^\infty\rho^n f^{n*}(x)dx, \qquad (A.15)$$

where $f^{n*}(x)$ is the n–fold convolution of the density of the distribution function of the remaining sizes $f(u)=\beta_1^{-1}[1-B(u)]$ with itself.

Thus the solutions of the system of the equations (A.9), (A.10) are given by the equalities (A.13) and (A.15).

The system (A.11), (A.12) is more complicated, but we do not have to solve it. We proceed as follows: let us multiply (A.12) by $\lambda(1-\rho)^{-1}[1-B(x)]$, integrate with respect to x from 0 to ∞, and add the resulting equation to (A.11). Then,

$$\frac{d}{du}\left[\delta_2(u)+\frac{\rho}{1-\rho}\bar\varphi_2(u)\right]=2\left[1+\lambda\int_0^\infty\varphi_1(y,u)dB(y)\right]\left[\delta_1(u)+\frac{\rho}{1-\rho}\bar\varphi_1(u)\right], \qquad (A.16)$$

where $\bar\varphi_2(u)$ is given by (A.5).

A similar procedure for the system (A.9), (A.10) leads to

$$\frac{d}{du}\left[\delta_1(u)+\frac{\rho}{1-\rho}\bar\varphi_1(u)\right]=\frac{1}{1-\rho},$$

whence it follows

$$\delta_1(u)+\frac{\rho}{1-\rho}\bar\varphi_1(u)=\frac{u}{1-\rho}, \qquad (A.17)$$

which, in virtue (A.3), yields $v_1(u)$.

Keeping (A.9) and (A.17) in mind, we rewrite the equation (A.16) in the form

$$\frac{d}{du}\left[\delta_2(u) + \frac{\rho}{1-\rho}\bar{\varphi}_2(u)\right] = \frac{2u}{1-\rho}\frac{d\delta_1(u)}{du}.$$

This last equation and (A.17) enable us to reduce (A.4) to the form

$$v_2(u) = \frac{2u^2}{(1-\rho)^2} - \frac{2}{1-\rho}\int_0^u \delta_1(x)dx,$$

where $\delta_1(x)$ is given by (A.15). Obviously, the variance is given by

$$\mathsf{Var}[V(u)] = \frac{u^2}{(1-\rho)^2} - \frac{2}{1-\rho}\int_0^u \delta_1(x)dx. \tag{A.18}$$

From the form of (A.18) it immediately follows that the coefficient of variation of the sojourn time distribution is less than unity. In other words the (conditional) distribution of the random variable $V(u)$ belongs to the class IFR (Increasing Failure Rate) probability distributions.

The formula (3.10) is obtained from (A.18) (see [10] for details).

As examples of the use of (A.18) we give below the exact expressions for $\mathsf{Var}[V(u)]$ obtained in closed form for the following distributions of job's sizes (type M, E_2 and H_2, respectively):

a)

$$B_1(x) = 1 - \mathrm{e}^{-\mu x},$$

b)

$$B_2(x) = 1 - \mathrm{e}^{-\mu x} - \mu x \mathrm{e}^{-\mu x},$$

c)

$$B_3(x) = \gamma(1 - \mathrm{e}^{-\mu_1 x}) + (1-\gamma)(1 - \mathrm{e}^{-\mu_2 x}).$$

We note that $\delta_1(u)$ in (A.18) depends on $B(x)$, and it is derived by means of inversion of (A.14) (see the equality (A.15)). We obtain after some algebra that in case a)

$$\mathsf{Var}[V(u)] = \frac{2\rho u}{\mu(1-\rho)^3} - \frac{2\rho}{\mu^2(1-\rho)^4}\left[1 - \mathrm{e}^{-\mu(1-\rho)u}\right]. \tag{A.19}$$

In cases b) and c) we have

$$\mathsf{Var}[V(u)] = -\frac{2hu}{1-\rho} - \frac{2}{(1-\rho)(y_1-y_2)}\left[\frac{hy_2 + \rho/(1-\rho)}{y_1}\left(1 - \mathrm{e}^{-y_1 u}\right)\right.$$
$$\left. - \frac{hy_1 + \rho/(1-\rho)}{y_2}\left(1 - \mathrm{e}^{-y_2 u}\right)\right]. \tag{A.20}$$

The terms of the right–hand side of (A.20) are:

for $B_2(x)$ — $\rho = 2\lambda/\mu$, $h = -3\rho/(2\mu(1-\rho)^2)$, $y_{1,2}$ are the roots of the quadratic equation

$$y^2 - (2\mu - \lambda)y + \mu^2 - 2\lambda\mu = 0;$$

for $B_3(x)$ — $\rho = \gamma\lambda/\mu_1 + (1-\gamma)\lambda/\mu_2$, $h = (\lambda - (y_1+y_2)\rho/(1-\rho))/(y_1y_2)$, $y_{1,2}$ are the roots of the quadratic equation

$$y^2 - (\mu_1 + \mu_2 - \lambda)y + \mu_1\mu_2 - \lambda\mu_1 + \lambda\mu_1\gamma - \lambda\mu_2\gamma = 0.$$

REFERENCES

1. Kleinrock L. Time-shared systems: a theoretical treatment. *J. Assoc. Comput. Mach.*, 1967, vol. 14, no. 2, pp. 242–251.
2. Kleinrock L. *Queueing Systems*. New-York: Wiley, 1976, vol. 2. Russian edition: Kleinrock L. *Computer Systems with Queues*. Moscow: Mir, 1979.
3. 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.
4. Yashkov S.F. Processor–sharing queues: some progress in analysis. *Queueing Systems*, 1987, vol. 2, no. 1, pp. 1–17.
5. Yashkova A.S., Yashkov S.F. Distribution of the virtual sojourn time in the M/G/1 procesor sharing queue. *Information Processes*, 2003, vol. 3, no. 2, pp. 128–137 (the journal is available via http://www.jip.ru/).
6. Kobayashi H., Konheim A. Queueing models of computer communications system analysis. *IEEE Trans. Commun.*, 1977, vol. 25, no. 1, pp. 2–28.
7. Jaiswal N.K. Performance evaluation studies for time–sharing computer systems. *Performance Evaluation,* 1982, vol. 2, no. 4, pp. 223–236.
8. 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). English edition: *J. of Soviet Mathematics*, 1992, vol. 58, no. 2, pp. 101–147.
9. Yashkov S.F. Some results of analyzing a probabilistic model of remote processing systems. *Autom. Control and Computer Sci.*, 1981, vol. 15, no. 4, pp. 1–8 (English edition of the Latvian journal *Avtom. i Vychislit. Tekhnika*, 1981, no. 4, pp. 3–11 by Allerton Press, USA).
10. Yashkov S.F. *Analysis of Queues in Computers* (in Russian with English Summary). Moscow: Radio i Svyaz, 1989 (review in *Mathematics and Computers in Simulation*, 1991, vol. 33, no. 2, pp. 177–178).
11. Schassberger R. A new approach to the M/G/1 processor–sharing queue. *Adv. Appl. Prob.*, 1984, vol. 16, no. 1, pp. 202–213.
12. Rege K.M. and Sengupta B. A decomposition theorem and related results for the discriminatory processor sharing queue. *Queueing Systems*, 1994, vol. 18, no. 3–4, pp. 333–351.
13. Whitt W. The M/G/1 processor–sharing queue with long and short jobs. *Unpublished manuscript*. 1998 (Sept.).
14. A.Ward and W.Whitt. Predicting response times in processor–sharing queues, in: *Analysis of Communication Networks: Call Centres, Traffic and Performance*, eds. D. R. McDonald and S.R.Turner (Fields Inst. Communications **28**), Providence: AMS, 2000, pp. 1–29.
15. Brandt A., Brandt M. A simple path relation for the sojourn times in G/G/1–PS system and its applications. ZIB–Report 03–18, Berlin, June 2003, pp. 1–14 (preprint Konrad–Zuse–Zentrum für Informationstechnik).
16. B.K.Asare and F.G.Foster. Conditional response times in the M/G/1 processor–sharing system, *J. Appl. Prob.*, 1983, vol. 20, no. 4, pp. 910–915.
17. Brandt A., Brandt M. On the sojourn times for many–queue head–of–the–line processor–sharing systems with permanent customers. *Math. Methods in Oper. Res.*, 1998, vol. 47, pp. 181–220.
18. Yashkov S.F., Yashkova A.S. The M/G/1 processor–sharing system: transient solutions. In: *Distributed Comput. Commun. Networks. Proc. 2nd Int. Conf.* (Tel–Aviv, Nov. 4-8, 1997). Moscow: Inst. for Info. Transm. Probl., 1997. pp. 261–272.
19. Takács L. *Introduction to the Theory of Queues.* New York: Oxford Univ. Press, 1962.
20. Sakata M., Noguchi S., Oizumi J. Analysis of a processor shared model for time sharing systems. In: *Proc. 2nd Hawaii Int. Conf. on System Sci.* Honolulu: Univ. of Hawaii, 1969, pp. 625–628.

21. Yashkov S.F, Yashkova A.S. Processor sharing queue: transient solutions. In: *Computer Sci. and Info. Technologies. Proc. 2nd Int. Conf.* (Yerevan, Aug. 17–22, 1999). Yerevan: National Acad. of Sci. of Armenia, 1999, pp. 99–103.

22. Zwart A.P., Boxma O.J. Sojourn time asymptotics in the M/G/1 processor sharing queue. *Queueing Systems*, 2000, vol. 35, pp. 141–166.

23. Villela D., Pradhan P., Rubenstein D. Provisioning servers in the application tier for e–commerce systems. In: *Proc. 12th IEEE Int. Workshop on Quality of Service* (Montreal, June, 2004), 2004. pp. 57–66.

24. Yashkov S.F. A note on asymptotic estimates of the sojourn time variance in the M/G/1 queue with processor–sharing. *Systems Analysis. Modelling. Simulation*, 1986, vol. 3, no. 3, pp. 267–269.

25. Yashkov S.F. Distribution of the conditional waiting time in a system with time-sharing. *Engineering Cybernetics*, 1977; vol. 15, no. 5, pp. 44–52 (English edition of the Russian journal *Izv. of the USSR Acad. of Sci. Tekhn. Kibernetika*, 1977, no. 5, pp. 88–94 by Scripta Publ. Co., USA).