Информационные процессы, Том 4, № 2, 2004, стр. 175–187.

THE TIME–DEPENDENT SOLUTION OF THE M/G/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 17, 2004

Abstract—This work deals with time–dependent analysis of a random process associated with the number of jobs and of their spent service times in the M/G/1 queueing system under the foreground–background (FBPS) processor sharing discipline. This discipline assumes that only the set of jobs with the least amount of attained service share the server in the pure processor sharing fashion. We derive the time–dependent distribution of the number of jobs each of which has an attained service time $a \leq y$ at time t in terms of the double transforms (with respect to space: the Laplace functional, and with respect to time: the Laplace transform) given the system is empty at time $t = 0$. In other words, it is obtained the non–stationary distribution of the random counting measure representing the transient state of the FBPS queue in all details. The method of the analysis is an extension and refinement of the principally new approach introduced and developed in the previous authors works (see, for example, [11]). We consider also some important special cases.

1. INTRODUCTION

The study of processor sharing queues is of considerable interest in modern queueing theory in connection with the mathematical modeling of the WEB servers in the INTERNET, time–sharing computer systems, asynchronous transfer mode in ISDN schemes, etc. During the last three decades considerable attention has been paid to the analysis of various processor sharing queueing models (the idea of processor sharing is originally due to Kleinrock (1967), an account of some important results can be found in Kleinrok [1, Ch. 4], and Yashkov [2]). One of the basic models is the M/G/1 queue with the foreground–background processor sharing (FBPS) discipline introduced also by Schrage [3]. This model has been studied by Kleinrock [1], Schassberger [4], and Yashkov [5], [6], among others (see also [2], [7] and references herein), but all of these previous investigations have dealt with stationary (ergodic) behaviour. (The steady–state sojourn time distribution have been known from the works [1] and [5], see Theorem 1 in Appendix.) An exception is provided only since 1988 when Yashkov [8] has obtained explicit analytical expressions for the Laplace transform (LT) of the generating function for the number of jobs at time t for the M/G/1—FBPS model by means of a non–trivial generalization of his previous solution for stationary queue–length distribution [6], [9]. Related transient solution for the M/G/1 queue with egalitarian processor sharing (EPS) can be found in Yashkov [2] (and in the references therein, also since 1988). Recent progress in this area is reflected in [10]. A modification of the derivation of the time–dependent queue–length distribution for the M/G/1—FBPS model is given in [11], the contribution of which is an improved proof with details under some relaxing the previous assumptions from [8], [12], [13].

The main purpose of this paper is to extend our result from [11] to the case where we no longer require to study only characteristics of the total number of jobs at time t as in [11] but instead we consider the number of jobs at time t, each of which has an attained service time $a \leq y$. Then the state of the FBPS

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

system can be viewed as a random counting measure (point process) which is characterized by its Laplace functional. In other words, we obtain the generalization of the same kind as it was made by Schassberger [4] who considered only the stationary distribution of such random counting measure and showed that the corresponding Laplace functional has the representation related to the generating function for the steady–state number of jobs in [6], [9], [12]. We derive the non–stationary distribution of the random counting measure for the M/G/1—FBPS model (announced partially in [13]) representing the transient state of the queue in all details. To do this, it is exploited the new method of analysis from [6], [8] (the method is completely different from Schassberger's one). Besides, we do not require that

(i) the distribution of the service times has a density,

(ii) the offered load is less than 1.

The rest of the paper is organized as follows. In §2 we describe the FBPS model, give a background to the analysis and prove the main theorem 2.1. In §3 our interest will focus on an extension of the main theorem to all half–axis and on some important corollaries. Some of them are obtained by means of the combination with theorems from Appendix. The final section contains few closing remarks.

2. STUDY OF RANDOM PROCESSES IN THE M/G/1—FBPS QUEUE

2.1. Model description and notations

A description of the FBPS discipline is as follows: 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 $n,\ 1 \le n < \infty$. 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.

We consider the M/G/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. random variables with a general distribution $B(x)$ $(B(0+) = 0,\ B(\infty) = 1)$ the first moment of which is $\beta_1 < \infty$. Let $\beta(s)$ be the LST of $B(x)$, i.e., $\beta(s) = \int_0^\infty \exp(-sx)dB(x)$. The number of jobs at time t in the system will be denoted as $L(t)$, the non–Markovian process $\{L(t) : t \in [0,\infty)\}$ is defined on a suitably chosen (as in [6]) probability space. The offered load will be denoted as $\rho = \lambda\beta_1$. We assume that the system is empty at time $t = 0$, although it is not difficult to extend our results to the case of various initial conditions.

Let $L(t, y)$ denote the number of jobs each of which has an attained (spent) service time $a \le y$ at time t. Then $L(t) \equiv L(t,\infty)$. The process $\left\{L(t, y) : t \ge 0, y \in \mathbf{R}^+\right\}$ can be also viewed as a random counting measure or point process [14, p.179], [15, Ch.5], [16] on the Borel sets of $[0,\infty)$. For each fixed t, $\{L(t,\cdot)\}$ have a discrete (atomic) distribution on $[0,\infty)$ herein the epochs of jumps correspond to the age of the groups of jobs (with the same spent service) and the values of jumps correspond to the numbers of jobs in the groups. We shall at first consider the process $\{L(t, y)\}$ during the first busy period. Let $\zeta = \inf(t > 0 : L(t,\infty) = 0)$. The indicator function of an event $(\cdot)$ is denoted as $\mathbf{1}_{(\cdot)}$.

Definition 2.1. The function

$$\mathcal{L}_1(f,t) \doteq \mathsf{E}\left[\exp\left(-\int_0^\infty f(y)L(t,dy)\right)\mathbf{1}_{(\zeta>t)}\,|\,L(0,\infty)=1\right] \tag{2.1}$$

is the Laplace functional of the process $\{L(t, y) : t \le \zeta, y \in \mathbf{R}^+\}$ over the first busy period given it is started by one job. Here f is any nonnegative measurable function on $[0,\infty)$. For brevity, we shall often omit the condition in the right–hand side (RHS) of formulae like (2.1) since it is also indicated by the subscript in the left–hand side.

The Laplace transform (with respect to t) of $\mathcal{L}_1(f,t)$ is given as

$$\tilde{\mathcal{L}}_1(f,s) \doteq \int_0^\infty \exp(-st)\mathcal{L}_1(f,t)\,dt, \quad \operatorname{Re} s > 0. \tag{2.2}$$

One of our main goals is to obtain $\tilde{\mathcal{L}}_1(f,s)$. This problem is solved in §2.3.

2.2. Outline of the method

In this subsection we give an outline of the method and the main ideas behind it. This forms a background to the analysis of the FBPS model. Let us consider the process $\{L(t,y):\, t \geq 0, y \in \mathbf{R}^+\}$ stopped at time ζ, i.e., $L(t,\infty) = 0$ for $t \geq \zeta$. Introduce the process $\{X(t)\}$ by

$$X(t) = \begin{cases} \inf(y \geq 0:\, L(t,\infty) - L(t,y) = 0) & \text{if } t < \zeta \\ X(\zeta^-) & \text{if } t \geq \zeta. \end{cases}$$

This process describes the maximal age over all jobs present at time t and it takes the value of the age for a job that has the lowest priority at time t in the first busy period (because of the FBPS rule, such job may be not single at some t). More formally, this cumulative process is introduced in [6],[17] as a continuous additive functional of a Markov process that describes the dynamics of $\{L(t,\cdot)\}$ together with additional coordinates which record the age of each job. ‘An inverse’ process to $\{X(t)\}$ is defined by

$$\tau(x) = \inf(t > 0:\, X(t) > x). \tag{2.3}$$

We use convention that $\inf\emptyset = +\infty$. Then $\tau(x)$ is right–continuous. Because of the FBPS discipline, the sample paths of $\{X(t)\}$ are continuous and non–decreasing functions. In view of this, the process $\{\tau(x)\}$ is right–continuous and strictly increasing. We note that $\tau(x)$ is finite for each finite x. Define now

$$M(x) = L(\tau(x),\infty). \tag{2.4}$$

Put also $M(x) = 0$ if $\tau(x) = \infty$. The new process $\{M(x)\}$ is obtained from $\{L(t,\infty)\}$ (together with its additional coordinates) by means of a random time change determined by (2.3). In other words, each realization of the process $\{L(t,\cdot)\}$ is endowed with its own clock, running individually with variable speed. The features of $\{M(x)\}$ and $\{\tau(x)\}$ are considered in [6], [17, Ch. 3] in the context of random processes in queues. In particular, it is known that, for each x, $\tau(x)$ is a stopping time, $X(\tau(x)) = x$, and the time–changed process $\{M(x)\}$ has the strong Markov property. Moreover, $\{M(x)\}$ is a nonhomogeneous Markov jump process. (Thus, by a random time change we simply understand an increasing net $\{\tau(x)\}$ of stopping times.)

In the queueing context, the transformation (2.4) means that we eliminate subbusy periods in which the jobs with the age $a < x$ are served, and, besides, we telescope the time scale, i.e., the time intervals during which the processor serves the jobs with the maximal age are compressed by a factor of n when the number of such jobs is equal to n. When the time scale is changed this way, the service rate is no longer a function depending on the number of jobs and their ages, but it is constant and equal to unit. However, the rate of the input process, when viewed through the new time scale, is the random process $\{\lambda M(x):\, x > 0\}$. For each $x < \infty$, $\tau(x)$ can be considered as an epoch (in old time scale) when the attained service of the lowest priority group of jobs reaches to x.

Remark 2.1. For each fixed x, the random variable $\tau(x) \wedge \zeta$ may be interpreted as the duration of some (sub)busy period during which the processor serves the jobs of the length $u \leq x$ completely, i.e., until they leave the system, and until the attained service of longer jobs (with the lengths $u > x$) reaches to the maximal attained service time x. For brevity, such busy period will be called an x–busy period.

We already saw that the process $\{M(x) : x > 0\}$ takes the values of the number of jobs with the maximal age x in the first busy period at the epochs $\{\tau(x)\}$. In virtue of this and Remark 2.1, we have

Remark 2.2. For each fixed $x > 0$, $M(x) : x > 0$ is related to, but not equal to, the number of jobs served during an x–busy period.

The above arguments enable us to reduce an investigation of complex queueing processes in the FBPS system to that of more simple and analytically tractable processes.

Remark 2.3. The idea of using the inverse of additive functional as a random time change in the process $\{L(t, \cdot)\}$ has a long history. It is hardly an uncommon one in probability theory. In context of the general theory of Markov processes, the random time change was first proposed by Volkonsky already in 1958 [18] and later was developed by many authors (see, e.g., the survey by Syski [19] or [20, Ch. 5]). In the queueing context, the idea of such sort was seemingly first developed in [6], [12] (see also [17, Chs. 2, 3]).

2.3. Distribution of point process on the first busy period

The main result of this paper is given in the following theorem. Its proof does not require smoothness assumptions on $B(\cdot)$.

Theorem 2.1. *The exact solution for* $\tilde{\mathcal{L}}_1(f, s)$ *is*

$$\tilde{\mathcal{L}}_1(f, s) = \lambda^{-1}\left[\exp\left(\lambda \int_0^\infty \left(z\frac{\partial Q(z, s, y)}{\partial z}\right)\bigg|_{z=\mathrm{e}^{-f(y)}} dy\right) - 1\right], \tag{2.5}$$

where function Q *is, for* $0 \le z \le 1$, $\operatorname{Re} s > 0$ *and* $x \ge 0$, *the unique solution of the functional equation*

$$Q(z, s, x) = \beta(s + \lambda - \lambda Q(z, s, x); x) + z(1 - B(x)) \exp(-x(s + \lambda - \lambda Q(z, s, x))). \tag{2.6}$$

Here $\beta(s; x) = \int_{0-}^{x} \exp(-su)\, dB(u)$.

Proof. **Step 1.** In virtue of the Fubini theorem, (2.2) is rewritten as

$$\tilde{\mathcal{L}}_1(f, s) = \mathsf{E}\left[\int_0^\zeta \exp(-st - I(f, t))\, dt\right],$$

where $I(f, t) = \int_0^\infty f(y)L(t, dy)$. Because $\tau(x)$ is a stopping time, we can introduce the following function

$$\tilde{\mathcal{L}}_1(f, s, x) = \mathsf{E}\left[\int_0^{\tau(x)\wedge\zeta} \exp(-st - I(f, t))\, dt\right]. \tag{2.7}$$

Note that the monotonous convergence yields

$$\tilde{\mathcal{L}}_1(f, s) = \lim_{x\to\infty} \tilde{\mathcal{L}}_1(f, s, x). \tag{2.8}$$

The equation for $\tilde{\mathcal{L}}_1(f, s, x)$ can be constructed by the following way. Suppose that Δ is infinitesimal, then

$$\tilde{\mathcal{L}}_1(f, s, x + \Delta) = \tilde{\mathcal{L}}_1(f, s, x) + \mathsf{E}\left[\int_{\tau(x)}^{\tau(x+\Delta)} \exp(-st - I(f, t))\mathbf{1}_{(\zeta>\tau(x))}\, dt\right]. \tag{2.9}$$

The last term in the RHS of (2.9) can be represented by the law of total probability as

$$\mathsf{E}\left[\int_{\tau(x)}^{\tau(x+\Delta)} \exp(-st - I(f, t))\mathbf{1}_{(\zeta>\tau(x))}\, dt\right] = \sum_{n=1}^{\infty} q_n(x)G_n(f, s, x, \Delta), \tag{2.10}$$

where $q_n(x) \doteq \mathsf{P}(L(\tau(x),\infty)=n)$ and

$$G_n(f,s,x,\Delta) \doteq \mathsf{E}\left[\int_{\tau(x)}^{\tau(x+\Delta)} \exp(-st - I(f,t))\,dt \,|\, L(\tau(x),\infty)=n\right]. \tag{2.11}$$

Step 2. The proof proceeds via preliminary lemma.

Lemma 2.1.

$$G_n(f,s,x,\Delta) = n\Delta\mathsf{E}\left[\mathrm{e}^{-s\tau(x)-f(x)M(x)} \,|\, M(x)=n\right] + \lambda\Delta n\mathsf{E}\left[\mathrm{e}^{-s\tau(x)-f(x)M(x)}\tilde{\mathcal{L}}_1(f,s,x) \,|\, M(x)=n\right] + o_n(\Delta), \tag{2.12}$$

where $M(x)$ is given by (2.4).

Proof. In view of our arguments from §2.2, the variable $\tau(x+\Delta)$ is composed of $\tau(x)+n\Delta$ and, possibly, of an x–busy period which is started by a new arrival in the segment $[\tau(x),\ \tau(x)+n\Delta]$. The first term in the RHS of (2.12) is the result of approximation of the integral in (2.11) over $[\tau(x), \tau(x)+n\Delta]$. The second term takes into account the possible interruption (i.e., an x–busy period) that occurs with probability $\lambda n\Delta$. We denote the number of jobs over this x–busy period as $\bar{L}(t,\cdot)$. Because of the FBPS discipline, the process $\left\{\bar{L}(t,\cdot)\right\}$ is an independent probabilistic copy of $\{L(t,\cdot)\}$ (the time is counted here from the start of the new x–busy period). Let $\bar{\tau}(x)$ and $\bar{\zeta}$ be defined for $\left\{\bar{L}(t,\cdot)\right\}$ as $\tau(x)$ and ζ, respectively, are defined for $\{L(t,\cdot)\}$ (see (2.3) and §2.1). Then, by Lebesgue's theorem for measure decomposition [14], the distribution of $\{L(t,\cdot)\}$ in $[\tau(x),\ \tau(x)+\bar{\tau}(x)\wedge\bar{\zeta})$ is the sum of the distribution of $\left\{\bar{L}(t-\tau(x),\cdot)\right\}$ and of a distribution that has single jump $M(x)$ at point x. Hence

$$\int_{\tau(x)}^{\tau(x)+\bar{\tau}(x)\wedge\bar{\zeta}} \mathrm{e}^{-st-I(f,t)}\,dt = \mathrm{e}^{-s\tau(x)-f(x)M(x)} \int_0^{\bar{\tau}(x)\wedge\bar{\zeta}} \mathrm{e}^{-st-\bar{I}(f,t)}\,dt,$$

where $\bar{I}(f,t) = \int_0^\infty f(y)\bar{L}(t,dy)$. This leads to the conditional expectation in the second term of the RHS of (2.12) in view of (2.7) and stochastic equivalence of $\left\{\bar{L}(t,\cdot)\right\}$ and $\{L(t,\cdot)\}$. □

If we now define

$$Q(z,s,x) \doteq \mathsf{E}[\mathrm{e}^{-s\tau(x)}z^{M(x)}],\ (\mathrm{Re}\,s>0,\ |z|\le 1), \tag{2.13}$$

then we obtain from (2.9), (2.10) and Lemma 2.1 the equation

$$\frac{\partial\tilde{\mathcal{L}}_1(f,s,x)}{\partial x} = \left(z\frac{\partial Q(z,s,x)}{\partial z}\right)\bigg|_{z=\mathrm{e}^{-f(x)}} [1+\lambda\tilde{\mathcal{L}}_1(f,s,x)] \tag{2.14}$$

by taking the limit as $\Delta\to 0$. We ought to take into account that $M(x)Q(z,s,x) = z(\partial Q(z,s,x)/\partial z)$ by virtue of (2.13). The initial condition is $\tilde{\mathcal{L}}_1(f,s,0)=0$. The solution of (2.14) is

$$\tilde{\mathcal{L}}_1(f,s,x) = \lambda^{-1}\left[\exp\left(\lambda\int_0^x \left(z\frac{\partial Q(z,s,y)}{\partial z}\right)\bigg|_{z=\mathrm{e}^{-f(y)}} dy\right) - 1\right],$$

from which it follows the assertion (2.5) in view of (2.8). □

Step 3. The last step is to show that the assertion (2.6) holds. For fixed $x,\ x>0$, let

$$H(u;x) = \begin{cases} B(u) & \text{if } u<x \\ 1 & \text{otherwise,} \end{cases}$$

then the LST of such truncated distribution is

$$h(s;x) = \int_0^x \mathrm{e}^{-su}\,dB(u) + \mathrm{e}^{-sx}(1-B(x)). \tag{2.15}$$

Note that $H(u;x)$ is the distribution of the random variable $B \wedge x$. We denote the number of jobs served during x–busy period by $N(x)$ and define the function

$$F(z,s,x) \doteq \mathsf{E}\left[\mathrm{e}^{-s\tau(x)}z^{N(x)}\right], \quad (\mathrm{Re}\, s > 0,\ |z| \le 1).$$

Taking into account Remark 2.1, we obtain that, if $H(u;x)$ is used instead of $B(u)$, $F(z,s,x)$ is found as the unique solution of the following functional equation for an M/G/1 queue

$$F(z,s,x) = zh(s+\lambda-\lambda F(z,s,x);x), \tag{2.16}$$

where x is viewed as a parameter. (For $x=\infty$ this equation gives the transform of joint distribution of the duration of the standard busy period and the number of jobs served during the busy period; it is the so–called Takacs equation [21, §1.3]. See also [22, p. 500] for an explicit expression of such joint distribution, but it is not in closed form.) Let u be the length of the job that initiates x–busy period. Taking into account Remark 2.2, we can see that

$$M(x) = \begin{cases} N(x)-1 & \text{if } u < x \\ N(x) & \text{if } u \ge x, \end{cases}$$

then this and (2.16) lead to assertion (2.6) in view of (2.13) and (2.15).

□

Remark 2.4 contains an additional comment to the step 3 of the proof of Th. 2.1.

Remark 2.4. In the new time scale x, the input process to the FBPS queue can be considered as the sum of a random number $M(x)$ of independent Poisson processes, each of rate λ. The process of such kind is closely related to a branching process in which the offsprings of a single ancestor are generated with rate λ, and they have the life lengths distributed as $B(\cdot)$. So, the process $\{M(x)\}$ coincides in distribution with such branching process of the number of particles, and $\tau(x) \wedge \zeta$ coincides in distribution with the sum of life lengths of all particles which fell on $[0,x]$. Thus, $Q(z,s,x)$ represents the joint distribution (in terms of double transform) of the number of particles and of their total age in the process with a single ancestor. The equation (2.6) can be interpreted as the decomposition of Q by the law of total probability for two events: the life length of the ancestor has expired at time $u \le x$ (the first term of the RHS of (2.6)), and the ancestor is alive at time x with probability $1-B(x)$ (the second term).

Remark 2.5. We can also construct partial–differential equation for $\{M(x)\}$, the solution of which is given by (2.6). Such less convenient approach is used in [6], [8], [9] for special case $s=0$.

Remark 2.6. The special case of (2.6) for $x \to \infty$ and $z=1$ gives the equation for the (standard) busy period distribution in terms of the LST $\pi(s) \doteq \mathsf{E}\left[\mathrm{e}^{-s\zeta}\right]$

$$\pi(s) = \beta(s+\lambda-\lambda\pi(s)). \tag{2.17}$$

It follows from [21, §1.3, Ths. 3, 4] that, when $\rho > 1$, (2.17) has two solutions as $s=0$: $\pi_1(0)=1$ and $\pi_2(0)=\pi(0)<1$ (we ought to choose the minimal one). As $s>0$, (2.17) has unique solution $\pi(s)$, herein $\pi(s)\to\pi(0)$ as $s\to 0$, hence Π, the busy period, is equal to infinity with probability $1-\pi(0)$. When $\rho \le 1$, (2.17) has unique solution $0<\pi(s)\le 1$ for each $s\ge 0$ and $\pi(s)\to 1$ as $s\to 0$, herein $\mathsf{E}[\Pi]=\infty$ if $\rho=1$ (when $\rho<1$, $\pi(0)=1$ and $\mathsf{E}[\Pi]<\infty$). Using argument of such kind, it is not very difficult to obtain the solution of (2.6) for any ρ. We do not require $Q(1,0,\infty)=1$, so Theorem 2.1 holds not only for the stability condition $\rho<1$ but even for $\rho\ge 1$.

Remark 2.7. A counterpart of Theorem 2.1 for the M/G/1—EPS queue can be obtained from the results of [17, §2.8] by similar extension; its representation resembles the formula (2.48) in [2].

3. POINT PROCESS ON ALL HALF–AXIS. SOME CONSEQUENCES

The well–known arguments from renewal theory allow us to connect (2.5) with the distribution of the process $\{L(t,y) : t \geq 0, y \in [0,\infty)\}$ on all half–axis. To this end, we ought to consider two renewal processes formed by points when the consecutive busy periods are initiated and finished, respectively. Let the whole FBPS system started initially empty.

Definition 3.1.

$$\mathcal{L}_0(f,t) \doteq \mathsf{E}\left[\exp\left(-\int_0^\infty f(y)L(t,dy)\right) \mid L(0,\infty)=0\right],$$

$$\tilde{\mathcal{L}}_0(f,s) = \int_0^\infty \exp(-st)\mathcal{L}_0(f,t)\,dt. \tag{3.1}$$

Theorem 3.1. *The exact solution for $\tilde{\mathcal{L}}_0(f,s)$ is*

$$\tilde{\mathcal{L}}_0(f,s) = [s+\lambda-\lambda\pi(s)]^{-1}[1+\lambda\tilde{\mathcal{L}}_1(f,s)],$$

where $\pi(s)$ is given by (2.17).

Proof. Similarly to the derivations of (3.13) or (2.107) from [23, Ch.1] or [17], respectively. We omit the details. □

Remark 3.1. Theorem 3.1 holds for an M/G/1 queue under any work–conserving discipline with the initial condition $\mathsf{P}(L(0,\cdot)=0)=1$. The discipline determines the form of function $\tilde{\mathcal{L}}_1(f,s)$. It is not very difficult to extend the result to the case $\mathsf{P}(L(0,\cdot)=n)=1$.

Our solution can be also generalized to the case of bulk arrivals. It will be changed only the initial condition and the offered load ρ. (It is done similarly to the study of stationary M/G/1—FBPS queue with bulk arrivals [6].)

We proceed to corollaries from Theorems 2.1 and 3.1.

Corollary 3.1. *For $\rho < 1$, the Laplace functional for the stationary process $\{L(t,y)\}$ is*

$$\mathcal{L}(f) = (1-\rho)\exp\left(\lambda\int_0^\infty \left(z\frac{\partial Q(z,0,y)}{\partial z}\bigg|_{z=\mathrm{e}^{-f(y)}} dy\right)\right),$$

where the function Q is given by (2.6) for the special case $s=0$.

Proof. Since the stationary distribution exists by Smith's ergodic theorem, then we obtain from Theorems 2.1 and 3.1 by using Abelian theorem for Laplace transforms

$$\mathcal{L}(f) = \lim_{t\to\infty}\tilde{\mathcal{L}}_0(f,s) = \lim_{s\downarrow 0}\, s\tilde{\mathcal{L}}_0(f,s),$$

and the result follows after routine algebra taking into account that

$$\lim_{s\downarrow 0}\, s(s+\lambda-\lambda\pi(s))^{-1} = 1-\rho$$

by virtue of L'Hospitale's rule. □

Remark 3.2. This stationary solution was obtained by Schassberger [4] who used another method — via a discrete time approximation of the FBPS model.

Function f can have, for example, the form $f(x) = \sum_{i=1}^{k} c_i \mathbf{1}_{x\in[a_i,b_i]}$ where the $c_i < \infty$ are real and $[a_i, b_i]$ are segments on time axis ($f(x)$ is a simple function in this case). If $f(x) = c = \mathrm{const}$, then we let $z = \mathrm{e}^{-c}$. For this special case, Theorems 2.1 and 3.1 give us the distribution of the number of jobs at time t [2, Th. 4.5 and (2.50)], [8, p. 43, 46], [11, Th. 2], [13].

Corollary 3.2. *For $f(x) = c = \mathrm{const}$, let $z = \mathrm{e}^{-c}$, $|z| \le 1$ then $\tilde{\mathcal{L}}_0(z, s)$ and $\tilde{\mathcal{L}}_1(z, s)$ are the LT's of the generating functions for the number of jobs at time t on all half–axis and on the first busy period, respectively. It holds*

$$\tilde{\mathcal{L}}_0(z, s) = [s + \lambda - \lambda\pi(s)]^{-1}[1 + \lambda\tilde{\mathcal{L}}_1(z, s)],$$

$$\tilde{\mathcal{L}}_1(z, s) = \lambda^{-1}\left[\exp\left(\lambda z \int_0^\infty \frac{\partial Q(z, s, y)}{\partial z}\, dy\right) - 1\right].$$

Proof. Straight arithmetic. □

Remark 3.3. This non–stationary solution was first obtained in [8] (see also [13]) under the assumption that $B(x)$ has a density; an improved proof with details and without the smoothness assumption on $B(\cdot)$ is contained in [11].

For comparison with the last corollary, we summarize few main solutions for $\tilde{\mathcal{L}}_1(z, s)$ for the FCFS and the EPS disciplines, which are available at present (the queue is empty at time $t = 0$).

Theorem 3.2. *For the M/G/1—FCFS queue,*

$$\tilde{\mathcal{L}}_1(z, s) = \frac{1 - \beta(s + \lambda - \lambda z)}{s + \lambda - \lambda z}\,\frac{z - \pi(s)}{1 - z^{-1}\beta(s + \lambda - \lambda z)}.$$

Proof. This is old result, see [23, Ch.1, (2.19)]. □

Theorem 3.3. *For the M/G/1—EPS queue,*

$$\tilde{\mathcal{L}}_1(z, s) = \frac{z(1 - \pi(s))}{s + \lambda(1 - z)(1 - \pi(s))}.$$

Proof. See Theorem 2.7 in [17] or Theorem 2.17 in [2]. □

Remark 3.4. Remark 2.6 remains in force for Theorem 3.3.

Some further consequences of Theorems 2.1 and 3.1 are following.

Corollary 3.3. *For $\rho < 1$, the generating function of the stationary queue–length distribution of the M/G/1—FBPS system is*

$$\mathsf{E}[z^L] = (1 - \rho)\exp\left(\lambda z \int_0^\infty \frac{\partial Q(z, 0, u)}{\partial z}\, du\right),$$

where function Q is given by (2.6) for the special case $s = 0$.

Proof. Similarly to the proof of Corollary 3.1, taking into account that $f(x) = c = \mathrm{const}$ and $z = \exp(-c)$. □

This corollary makes it possible to compute first n moments of L. The first two moments which are found in [6] are:

$$\mathsf{E}[L] = \lambda \int_0^\infty m_2(x)\, dx, \tag{3.2}$$

$$\mathsf{Var}[L] = \lambda \int_0^\infty m_3(x)\,dx. \tag{3.3}$$

In (3.2) and (3.3), $m_2(x)$ and $m_3(x)$ are given by the formulas

$$m_2(x) = (1-\rho(x))^{-1}\lambda^2 m_1^2(x)h_2(x) + 2\lambda x m_1^2(x) + m_1(x), \tag{3.4}$$

$$\begin{aligned} m_3(x) = (1-\rho(x))^{-1}[\lambda^3 m_1^3(x)h_3(x) + 3\lambda^2 m_1(x)(m_2(x)-m_1(x))h_2(x)] \\ + 3\lambda^2x^2m_1^3(x) + 3\lambda x(m_2(x)-m_1(x))m_1(x) + 3m_2(x) - 2m_1(x). \end{aligned} \tag{3.5}$$

Here $m_1(x) = (1-B(x))/(1-\rho(x))$, $\rho(x) = \lambda h_1(x)$, where $h_1(x)$ is given by formula

$$h_j(x) = j\int_0^x y^{j-1}(1-B(y))\,dy, \quad j = 1,2,\dots. \tag{3.6}$$

It is the jth moment of (2.15). We note that $\lambda\int_0^\infty m_1(x)\,dx = -\ln(1-\rho)$.

Corollary 3.4. *Let* $\mathsf{P}_{00}(t) \doteq \mathsf{P}(L(t)=0|L(0)=0)$. *This probability has the LT w.r.t t*

$$p_{00}(s) = [s+\lambda-\lambda\pi(s)]^{-1}.$$

Proof. Immediately from Corollary 3.2, setting $z=0$. □

Let T_c be the relaxation time of the M/G/1—FBPS system. This quantity determines the rate of convergence to the steady–state. The relaxation time of a queue is usually defined to be the inverse of the exponent which dominates the asymptotic behaviour in the long run of the queueing characteristics, e.g., of $\mathsf{P}_{00}(t)$.

Corollary 3.5. *Let* $B(x)$ *has a light tail (i.e.,* $\beta(-s) < \infty,\ s > 0$*) and* $s_b < 0$ *be the abscissa of convergence of* $\beta(s)$. *Suppose also that* $\beta(s) \uparrow \infty$ *as* $s \downarrow s_b$. *Then* $T_c = -s_0^{-1}$, *where* s_0 *is the singularity with the largest real part of* $p_{00}(s)$ *determined by Corollary 3.4 (apart from a pole at* $s=0$*).*

Proof. Similarly to the comments for Theorem 2.19 in [2, p.120]. The first proof is due to Blanc and van Doorn [24]. □

Remark 3.5. Corollaries 3.4 and 3.5 remain in force for an M/G/1 system under any work–conserving discipline with the initial condition $\mathsf{P}(L(0)=0)=1$.

Remark 3.6. Significance of Theorems 2.1, 3.1, 3.3 grows in the light of the following facts.

(i) The discipline FBPS stochastically minimizes $L(t)$ for each t in the classes DFR (decreasing failure rate) and IMRL (increasing mean residual life) distributions of $B(x)$ in the work–conserving GI/G/1 queue when the remaining lengths of jobs are unknown [5], [25] (see Appendix for the corresponding theorems). For example, $B(x) \in DFR$ if and only if $\log(1-B(x))$ is convex. The FBPS coincides with the optimal strategy (policy) in such cases [25, Th.1].

(ii) The distributions of job lengths (for the current INTERNET state) belong, as a rule, to the heavy–tailed distributions ($\beta(-s)=\infty,\ s>0$) which studied earlier in insurance mathematics [15]. The typical example of an acceptable approximation for the file size is a family of the Pareto distributions.

(iii) The (standard) Pareto distribution has linearly decreasing failure rate.

Corollary 3.6. *If* $B(x) \in$ *DFR or IFR (see Appendix), then (3.2) reduces to*

$$\mathsf{E}[L] = \int_0^\infty (1-\rho(u))\int_{\rho(u)}^\rho \frac{\mu(x)}{(1-\rho(x))^2}d\rho(x)du,$$

where $\mu(x)$ *is given in Definition 1 in Appendix.*

Theorem 3.4. *If $B(x) \in$ DFR with monotone $\mu(x)$, then it holds for the queue GI/G/1 under the FBPS*

$$\mathsf{P}(L(t)_{FBPS} = k) \le \mathsf{P}(L_{FCFS}(t) = k) \quad w.p.1,$$

where $\mathsf{P}(L_{FCFS}(t) = k)$ are the time–dependent state probabilities of the queue GI/M/1 with FCFS.

Proof. Apply Theorem 3 from Appendix with trivial manipulations. □

Corollary 3.7. *It holds the following estimate for the steady–state M/G/1—FBPS in the case $B(x) \in$ DFR*

$$\mathsf{P}(L_{FBPS} \ge k) \le \rho^k, \quad k = 1, 2, \dots$$

This corollary can be also deduced from Corollary 3.3. However using Theorem 3.4 above or Theorem 2 from Appendix gives us the result almost immediately, taking into account the generating function of the number in the M/M/1—FCFS: $\mathsf{E}[z^L] = (1-\rho)/(1-\rho z)$. In view of our exact result (see Corollary 3.3), we shall not obtain the sharper bounds.

Example. This is only small part of the set of the examples from [17], data below are also contained in [6]. Consider the queue M/G/1 in the stedy–state with $\lambda = 0.8$. For the 2-stage hyperexponential distribution (of type $H_2 \in DFR$) $B(x) = 0.6(1 - e^{-3x}) + 0.4(1 - e^{-0.5x})$ with the moments $\beta_1 = 1$, $\beta_2 = 3.33$, $\beta_3 = 19.33$, we have that $\mathsf{E}[L]_{FBPS} = 3.57$, $\mathsf{E}[L]_{EPS} = 4.00$, $\mathsf{E}[L]_{FCFS} = 6.13$ and $\mathsf{Var}[L]_{FBPS} = 14.98$, $\mathsf{Var}[L]_{EPS} = 20.00$, $\mathsf{Var}[L]_{FCFS} = 52.45$. The results have been calculated from (3.2 and (3.3) for the FBPS case (the other results have been obtained from the well–known formulas). For the 2-stage Erlang distribution (of type $E_2 \in IFR$) $B(x) = 1 - e^{-2x} - 2xe^{-2x}$ with the moments $\beta_1 = 1$, $\beta_2 = 1.5$, $\beta_3 = 3$, we have that $\mathsf{E}[L]_{FBPS} = 4.43$, $\mathsf{E}[L]_{EPS} = 4.00$, $\mathsf{E}[L]_{FCFS} = 3.20$ and $\mathsf{Var}[L]_{FBPS} = 26.08$, $\mathsf{Var}[L]_{EPS} = 20.00$, $\mathsf{Var}[L]_{FCFS} = 11.80$. These data illustrate some properties of the FBPS discipline.

CONCLUSION

Our theorems and their corollaries provide new powerful tools for the deep study of time–dependent performance measures of the M/G/1—FBPS queue (and the M/G/1—EPS queue, too). These results open also the ways for investigation of the transient ($\rho > 1$) and null–recurrent ($\rho = 1$) behaviour of these basic models for predicting delays in the WEB servers. To efficiently apply the results, it can be also useful using the asymptotic methods of analysis (as in [26], [27]), developed only for stationary performance measures, or the methods of numerical inverting the Laplace transforms (for example, as in [28]).

APPENDIX

Here we briefly describe some basic results for the M/G/1—FBPS queue published in the papers [5], [25], which are difficult to access for English–language readers. Arrivals occur according to Poisson process with rate λ, the distribution of the job's lengths $B(x)$ ($B(0+) = 0$) has the LST $\beta(s)$, $\beta_1 = \int_0^\infty (1 - B(x))\,dx < \infty$, $\rho = \lambda\beta_1$. Let $V(u)$ be the stationary sojourn time for a tagged job which has a length u at the arrival time and $v(s,u) = \mathsf{E}[\mathrm{e}^{-sV(u)}]$.

Theorem 1. *(Yashkov [5], 1978). If $\rho < 1$ and $B(x)$ is absolutely continuous, then*

$$v(s,u) = w(s + \lambda - \lambda\pi(s,u), u)e^{-u(s+\lambda-\lambda\pi(s,u))},$$

where

$$w(s,u) = \frac{s(1-\rho(u))}{s - \lambda + \lambda h(s,u)},$$

$\rho(u) = \lambda h_1(u)$, $\pi(s,u) = h(s + \lambda - \lambda\pi(s,u), u)$. Here $h(s,u)$ and $h_j(u)$ are given by (2.15) and (3.6), respectively.

Comments to the proof. The proof in [5] uses the method of delay cycle analysis with additional considerations of main properties of the backlog process $U_x(t)$ (a unfinished work, truncated on level x) in work–conserving systems. □

Corollary 1. *(Yashkov [5], 1978).*

$$v_1(u) = \mathsf{E}[V(u)] = \frac{\lambda h_2(u)}{2(1-\rho(u))^2} + \frac{u}{1-\rho(u)},$$

$$\mathsf{Var}[V(u)] = \frac{\lambda h_3(u)}{3(1-\rho(u))^3} + \frac{\lambda u h_2(u)}{(1-\rho(u))^3} + \frac{3(\lambda h_2(u))^2}{4(1-\rho(u))^4},$$

$$\mathsf{E}[V] = \int_0^\infty v_1(u)\,dB(u) = \int_0^\infty (1-B(x))\,dv_1(x). \tag{A.1}$$

Remark 1. (Yashkov [5], 1978). It follows from (A.1) and Little's formula that

$$\mathsf{E}[L] = \lambda \int_0^\infty \left[\frac{\lambda \int_0^x u(1-B(u))\,du}{(1-\rho(x))^2} + \frac{x}{1-\rho(x)}\right] dB(x). \tag{A.2}$$

Eqs. (A.2) and (3.2) are equivalent.

Definition 1. Let $B(x)$ be absolutely continuous and $\mu(x) = dB(x)/(1-B(x))$. If $\mu(x)$ is decreasing (increasing) in x, then $B(x) \in$ DFR (IFR) distribution.

Theorem 2. *(Yashkov [5], 1978). If $B(x) \in$ DFR (IFR) with monotone $\mu(x)$, then* $\mathsf{E}[L]$ *and* $\mathsf{E}[V]$ *in the M/G/1—FBPS queue is minimal (maximal) over all possible work–conserving disciplines that do not take advantage of precise knowledge about the lengths of jobs.*

Comments to the proof. The main idea is to associate with each job the priority index

$$J(a) = \sup_{x \geq a} \frac{\int_a^x dB(u)}{\int_a^x (1-B(u))du}, \tag{A.3}$$

which is obtained from a discrete–time version of the model. Then the optimal order of service (which minimizes $\mathsf{E}[L]$ and $\mathsf{E}[V]$) is provided by giving the preemptive–resume priority to the jobs with maximum $J(a)$. If there $n > 0$ jobs with the same maximal $J(a)$, then such jobs must be serviced simultaneously with the rate $1/n$. It is a simple consequence to applying the representation of $B(x)$ in the form

$$B(x) = 1 - e^{-\int_0^x \mu(u)du}.$$

When $\mu(a)$ is decreases (increases) monotone on a, then (A.3) set up the FBPS (FCFS) discipline. If $\mu(a) = \text{const}$, then $\mathsf{E}[L]$ coincides with the same performance measure for the queue M/M/1 with any work–conserving discipline which does not use information about remaining lengths of jobs (e.g., the FCFS). □

Definition 2. If $(1-B(t))^{-1}\int_t^\infty (1-B(y))dy$ is increasing (decreasing) in t, then $B(x) \in$ IMRL (DMFL) distribution.

Theorem 3. *(Yashkov [25], 1991). The FBPS (FCFS) stochastically minimizes (maximizes) $L(t)$ at each t in the class of IMRL (DMRL) distributions of $B(x)$ among all work–conserving disciplines in the system GI/G/1, that do not take advantage of precise knowledge about the lengths of jobs.*

Comments to the proof. It is used path by path analysis of the FBPS queue via some stochastic differetial equation, taking into account that:

(i) the FBPS minimizes truncated unfinished work $U_x(t)$ in the case of IMRL distribution of $B(x)$ (it is some by–product in the proof of Theorem 1),

(ii) definition 2,

(iii) formal integration by parts,

(iv) interchanging the orders of taking the limit and of taking expectation in the corresponding stochastic integrals. □

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. 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.
3. Schrage L. The Queue M/G/1 with Feedback to Lower Priority Queues. *Manag. Sci.*, 1967, vol. 13, no. 7, pp. 466–474.
4. Schassberger R. Steady–state distribution of spent service times present in the M/G/1 foreground–background processor sharing queue. *J. Appl. Prob.*, 1988, vol. 25, no. 1, pp. 194–203.
5. 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).
6. Yashkov S.F. Analysis of a system with priority–based processor–sharing. *Avtom. i Vychisl. Tekhn.*, 1984, no. 3, pp. 29–38 (in Russian). Engl. edition: *Autom. Contr. and Comput. Sci.*, 1984, vol. 18, no. 3, pp. 27–36.
7. Aalto S., Ayesta U., Nyberg–Oksanen E. Two–level processor–sharing scheduling disciplines: mean delay analysis. In: *ACM SIGMETRICS/Performance'04*. New York, 2004.
8. Yashkov S.F. Non–stationary characteristics of the foreground–background processor sharing system (in Russian). In: *Proc. Int. Seminar on Teletraffic Theory and Computer Modelling* (Sofia, March 21–26, 1988). Sofia: Inst. of Math., 1988, pp. 41–46.
9. Yashkov S.F. Distribution of number of jobs in a feedback processor–sharing queue. In: *Systems Analysis and Simulation 1985. Proc. 2nd Int. Symp.* (Berlin, Aug. 26–31, 1985). Eds. Sydow A. et al. Berlin: Akademie, 1985, vol. 1 (Ser. Math. Research, Bd. 27), pp. 464–468.
10. 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).
11. 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.
12. Yashkov S.F. New application of random time change to analysis of processor–sharing queues. In: *4–th Int. Vilnius Conf. on Prob. Theory and Math. Statistics*. Vilnius: Inst. of Math., 1985, vol. 4, pp. 343–345.
13. Yashkov S.F. The foreground–background processor sharing queue: some developments in analysis. *5–th Int. Vilnius Conf. on Prob. Theory and Math. Statistics.*. Vilnius: Inst. of Math., 1989, vol. 2, pp. 236–237.
14. Bulinski A.V., Shiryaev A.N. *The Theory of Stochastic Processes.* Moscow: Fizmatlit, 2003 (in Russian).
15. Embrechts P., Kluppelberg C., Mikosch T. *Modelling Extremal Events for Insurance and Finance.* Heidelberg: Springer, 1997.
16. Brandt A., Franken P., Lisek B. Stationary Stochastic Models. Berlin: Akademie, 1990.
17. Yashkov S.F. *Analysis of Queues in Computers*. Moscow: Radio i Svyaz, 1989 (in Russian).

18. Volkonski V.A. Random substitution in strong Markov processes. *Teoriya Veroytn. i eye primen.*, 1958, vol. 3, no. 3, pp. 332–350. Engl. edition: *Theor. Prob. Appl.*, 1958, vol.3, no. 3, pp. 310–325.

19. Sysky R. Markov functionals in teletraffic theory. In: *Teletraffic Analysis and Computer Perform. Eval.*. Eds. Boxma O., Tijms H. Amsterdam: North Holland, 1986, pp. 303–317.

20. Blumenthal R., Getoor R. *Markov Processes and Potential Theory.* New York: Academic Press, 1968.

21. Takacs L. *Introduction to the Theory of Queues*. New–York: Oxford Univ. Press, 1962.

22. Cooper R.B. Queueing theory. Ch.10 in: *Handbooks in Oper. Res. and Manag. Sci.* vol. 2: Stochastic Models. Eds. Heyman D.P., Sobel M.J. New York: Elsevier, 1990, pp. 469–518.

23. Jaiswal N.K. *Priority Queues*. New York: Academic Press, 1968. Russian edition: *Queues with Priorities.* Ed. Kalashnikov V.V. Moscow: Mir, 1973.

24. Blank J.P.C., van Doorn E.A. Relaxation times for queueing systems. in: *Mathematics and Computer Science.* Eds. Hazewinkel M., Lenstra J.K. Amsterdam: North Holland, 1986, pp. 139–162.

25. Yashkov S.F. On optimality of the foreground–background processor sharing discipline. In: *Proc. 16–th All–Union School–Seminar on Computer Networks.* Moscow–Vinnitsa: Sci. Cybern. Council, 1991, pt. 3, pp. 103–107 (in Russian).

26. Jelenkovich P., Momchilovich P. Resource sharing with subexponential distributions. In: *Proc. IEEE Infocom'2002*. New York, 2002, pp. 1316–1325.

27. Borst S.C, Boxma O.J., Nunez–Queija R. Heavy tails: the effect of the service discipline. In: *Computer Performance Evaluation — Modelling Techniques and Tools. Proc. 12th Int. Conf. (London, Apr. 2002).* Eds. Field T. et al. Berlin: Springer, pp. 1–30.

28. Abate J., Whitt W. The Fourier–series method for inverting transforms of probability distributions. *Queueing Systems*, 1992, vol. 10, no. 1, pp. 5–88.

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