ISSN 0001-4346, Mathematical Notes, 2015, Vol. 97, No. 6, pp. 821–830. © Pleiades Publishing, Ltd., 2015.
Original Russian Text © S. Zh. Aibatov, 2015, published in Matematicheskie Zametki, 2015, Vol. 97, No. 6, pp. 803–814.

Ergodic Theorem for a Queue with Unreliable Server

S. Zh. Aibatov*

Lomonosov Moscow State University, Moscow, Russia

Received May 16, 2014; in final form, October 26, 2014

Abstract—A single server queue with unreliable server is considered. The server breakdown probability depends on the state of a random medium described by a Markov chain. An ergodic theorem is proved for this system.

DOI: 10.1134/S0001434615050181

Keywords: *regenerative flow, unreliable server, ergodicity, limit theorems, Markov chain, random medium.*

1. INTRODUCTION

Single server queues with unreliable server have been studied by numerous authors. This is largely due to a wide range of applications in highly diverse fields such as computer systems, telecommunications, transportation systems, airports, etc. The paper [1], where a model of queue with unreliable server was introduced, was one of the first papers on the topic. One should also mention the paper [2], where various types of interruptions are considered and the notion of *completion time*, which can be used to reduce a system with interruptions to an $M/G/1/\infty$ system, is introduced. The papers [3] and [4] deal with the topic as well. The paper [5] provides a description of existing results and a vast bibliography.

It is assumed in the model considered in the present paper that the server breakdown and resumption are caused by some external factors independent of the input flow and the in-service time. The process affecting the server operation is an ergodic Markov chain. One peculiar feature of this model is that the regeneration times and the times between interruptions are, in general, dependent, which distinguishes the present paper from those cited above.

The aim of our study is to find conditions for the existence of a proper limit distribution of the virtual waiting time in the system. The proof of the theorem in the present paper is largely based on the results of [6] as well as on the properties of regenerative flows.

The paper is organized as follows. The mathematical model is described in Sec. 2. Section 3 presents the proof of the ergodic theorem. Examples of applications of this theorem are given in Sec. 4.

2. DESCRIPTION OF THE MODEL

Consider a Reg $/G/1/\infty$ queue, where the symbol Reg indicates that the input flow is regenerative. Let us give a definition of regenerative flow following [7].

Definition 1. A stochastic process $\{A(t),\, t \geq 0\}$ with left continuous nondecreasing trajectories on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$ is called a *regenerative flow* if there exists an increasing sequence $\{\theta_j,\, j \geq 0\}$, $\theta_0 = 0$, of random variables such that

$$\{\kappa_j\}_{j=1}^{\infty} = \{\theta_j - \theta_{j-1},\, A(\theta_{j-1}+t) - A(\theta_{j-1}),\, t \in [0, \theta_j - \theta_{j-1})\}_{j=1}^{\infty}$$

is a sequence of independent identically distributed random elements. Further, we assume that a filtration $\{\mathscr{F}_t, t \geq 0\}$, $\mathscr{F}_t \subseteq \mathscr{F}$, is given, the process $A(t)$ is measurable, and each θ_j, $j = 1, 2, \ldots$, is a Markov moment with respect to $\{\mathscr{F}_t,\, t \geq 0\}$.

*E-mail: aybatov.serik@gmail.com

The sequence $\{\tau_i = \theta_i - \theta_{i-1}\}_{i=1}^{\infty}$ consists of independent identically distributed random variables, which are called the *regeneration times*, and $\xi_i = A(\theta_i) - A(\theta_{i-1})$ is the number of units arriving in the ith regeneration period. We assume that $\mathsf{E}\xi_i < \infty$ and $\mathsf{E}\tau_i < \infty$, so that, with probability 1, there exists a limit

$$\lim_{t\to\infty} \frac{A(t)}{t} = \frac{\mathsf{E}\xi_1}{\mathsf{E}\tau_1} = \lambda \qquad \text{(a.s.)},$$

which is called the *arrival rate*. A considerable part of flows customarily used in queuing theory are regenerative. These include doubly stochastic Poisson flows with random arrival rate representing a regenerative stochastic process (e.g., see [8]), semi-Markov queues (e.g., see [9]), Markov modulated queues (e.g., see [7]) and many others. Regenerative flows have a number of useful properties simplifying the analysis of systems whose input flows are regenerative. Many of these properties, which are used in what follows, were described and proved in [7].

Request service times are given by a sequence $\{\eta_i\}_{i=1}^{\infty}$ of independent identically distributed random variables independent of the input flow. Set $B(x) = \mathsf{P}\{\eta_i \le x\}$ and $b = \mathsf{E}\eta_1 < \infty$.

We need the following assumption, which ensures that the processes in question are regenerative (e.g., see [10]).

Condition 1. Let $\{t_n\}_{n=1}^{\infty}$ be the request arrival times. Then

$$\mathsf{P}\{\xi_1 = 0,\, \tau_1 > 0\} + \mathsf{P}\{\xi_1 = 1,\, \tau_1 - t_1 > \eta_1\} > 0. \tag{1}$$

The total service time for all request arriving on $[0, t)$ will be denoted by

$$X(t) = \sum_{j=1}^{A(t)} \eta_j.$$

Then $X(t)$ is a regenerative flow as well, with regeneration points $\{\theta_i\}$.

The server can break down, its breakdowns and time intervals between resumptions being related to a stochastic process $U(t)$ independent of $A(t)$ and $\{\eta_i\}$. We assume that $U(t)$ is an ergodic Markov chain with state space $\mathbb{E} = (0, 1, 2, \dots)$ and infinitesimal matrix $Q = (q_{ij})$.

Now let us describe how the process $U(t)$ affects the server. Whenever $U(t)$ switches into a state $i \in \mathbb{E}$, the server breaks down with probability α_i if it is up and resumes operation with probability β_i if it is down. Thus, the server breakdowns and resumptions can only occur at the switching times of the process $U(t)$.

This model can describe various situations. For example, if $U(t)$ is the number of requests in an $M/M/1$ system and

$$\alpha_1 = 1, \qquad \beta_0 = 1, \qquad \alpha_i = 0, \quad i \ne 1, \qquad \beta_i = 0, \quad i \ne 0,$$

then we actually deal with an absolute priority queue. There are requests of two types, $A(t)$ being the input flow of requests of the first type, while requests of the second type arrive in accordance with a Poisson flow and have exponentially distributed service time. Requests of second type have absolute priority, so that a request of the first type can only be serviced if there are no requests of the second type in the system.

Another situation occurs if the server being up or down depends on random external factors. The set of states $U(t)$ is represented as the union $E = E_+ \cup E_-$ of two sets, and

$$\begin{cases} \alpha_i > 0, \\ \beta_i = 0 \end{cases} \quad \text{if } i \in E_- \qquad \text{and} \qquad \begin{cases} \alpha_i = 0, \\ \beta_i > 0 \end{cases} \quad \text{if } i \in E_+.$$

Thus, the device can break down for states in E_- and resume operation for states in E_+.

We assume that the following condition holds.

Condition 2. The process $U(t)$ has states i_0 and i_1 such that $\alpha_{i_0} > 0$ and $\beta_{i_1} > 0$.

Without loss of generality, we set $i_1 = 0$.

Since all states are connected with each other, it follows that the time intervals on which the server is up (or down) are finite with probability 1 in view of Condition 2.

If the server breaks down when servicing a request, then servicing is continued at the resumption instant exactly from the point where it has been interrupted, so that the total uptime used to service the request has the distribution function $B(x)$.

Our model has three features distinguishing it from the one considered in [2]. In [2], the input flow is Poisson (i.e., an $M/G/1/\infty$ system is considered), and the up- and downtime are independent random variables. Moreover, it is assumed that the uptime has exponential distribution and that the server can only break down when servicing a request. These conditions do not hold in our model in general. Note also that the up- and downtime intervals are independent if, say, there exist j_0 and j_1 such that

$$\begin{aligned} &\alpha_{j_0} > 0, \qquad \alpha_j = 0 \quad \text{for} \quad j \neq j_0, \\ &\beta_{j_1} > 0, \qquad \beta_j = 0 \quad \text{for} \quad j \neq j_1; \end{aligned}$$

i.e., the server can break down or resume operation only for a single state $U(t)$.

We define a random medium for the system by using the process $N(t) = \{e(t), U(t)\}$, where $e(t) = 1$ is the server is up at time t and $e(t) = 0$ otherwise. Obviously, $N(t)$ is a continuous-time Markov chain with state space $\{0,1\} \times \mathbb{E}$, and the entries of its infinitesimal matrix have the form (for $i \neq j$)

$$\begin{aligned} &q_{(0,i)(0,j)} = q_{ij}(1-\beta_j), \qquad && q_{(0,i)(1,j)} = q_{ij}\beta_j, \\ &q_{(1,i)(0,j)} = q_{ij}\alpha_j, && q_{(1,i)(1,j)} = q_{ij}(1-\alpha_j), \\ &q_{(k,i)(k,i)} = q_{ii}, && q_{(k,i)(1-k,i)} = 0. \end{aligned} \tag{2}$$

The ergodicity of $U(t)$ implies that of $N(t)$. We set

$$\begin{gathered} p_i^0 = \lim_{t\to\infty} \mathsf{P}(e(t) = 0,\, U(t) = i), \qquad p_i^1 = \lim_{t\to\infty} \mathsf{P}(e(t) = 1,\, U(t) = i), \\ \pi = \lim_{t\to\infty} \mathsf{P}(e(t) = 1) = \sum_{i\in\mathbb{E}} p_i^1. \end{gathered} \tag{3}$$

The steady-state probabilities p_i^0, p_i^1 are a solution of the system

$$p_i^0 = \sum_{j\in\mathbb{E}} (p_j^0 p_{(0,j)(0,i)} + p_j^1 p_{(1,j)(0,i)}), \qquad p_i^1 = \sum_{j\in\mathbb{E}} (p_j^0 p_{(0,j)(1,i)} + p_j^1 p_{(1,j)(1,i)}),$$

where

$$p_{(k,j)(l,i)} = \frac{q_{(k,j)(l,i)}}{-q_{(k,j)(k,j)}}, \qquad k,l \in \{0,1\}, \quad i,j \in \mathbb{E},$$

and $q_{(k,j)(l,i)}$ can be found from (2).

3. ERGODIC THEOREM

Consider the process $W(t)$ representing the total remaining service time for the requests that are in the system at time t assuming that the server does not break down after time t.

Theorem 1. *Let $U(t)$ be an ergodic Markov chain, and let Conditions 1 and 2 be satisfied. Set $\rho = \lambda b/\pi$. Then*

- $W(t) \xrightarrow[t\to\infty]{a.s.} \infty$ *if* $\rho > 1$.
- $W(t) \xrightarrow[t\to\infty]{\mathsf{P}} \infty$ *if* $\rho = 1$.

- *For $\rho < 1$ and for an arbitrary initial state $W(0)$, there exists a limit*

$$\lim_{t\to\infty} \mathsf{P}(W(t) \le x) = F(x),$$

and $F(x)$ is a distribution function independent of $W(0)$.

Proof. Let us introduce the stochastic process

$$Y(t) = \int_0^t 1(e(s) = 1)\, ds,$$

which is the total server uptime on the interval $[0, t)$. Then (e.g., see [11])

$$W(t) = \sup_{0\le s\le t} [W(0) + Z(t),\ Z(t) - Z(s)], \tag{4}$$

where $Z(t) = X(t) - Y(t)$.

Note that $Y(t)$ has nondecreasing trajectories and is a regenerative flow. Its regeneration times form a subsequence $\{T_i\}_{i=1}^{\infty}$ of $\{\theta_i\}_{i=1}^{\infty}$ such that

$$e(T_i + 0) = 1, \qquad U(T_i + 0) = 0.$$

Since $X(t)$ and $Y(t)$ are independent, we see that the T_i are also regeneration times of $X(t)$. This follows from the properties of regenerative flows (independent sifting) [7]. Without loss in generality, we assume that $\mathsf{P}(U(0) = 0,\ e(0) = 1) = 1$. Then the sequence $\{\tau_k' = T_k - T_{k-1}\}_{k=1}^{\infty}$ consists of independent identically distributed random variables.

Lemma 1. *If Condition 2 is satisfied, then $\mathsf{E}\tau_k' < \infty$.*

Proof. Set

$$\nu_k = \min\{j > \nu_{k-1} : U(\theta_j) = 0,\ e(\theta_j) = 1\}, \qquad \nu_0 = 0,$$

so that $T_k = \theta_{\nu_k}$. Then $\{\nu_k - \nu_{k-1}\}_{k=1}^{\infty}$ is a sequence of independent identically distributed random variables, and $\{\nu_k\}_{k=0}^{\infty}$ is a regeneration process. By the Blackwell theorem, the regeneration function

$$h(j) = \sum_{k=1}^{\infty} \mathsf{P}(\nu_k = j) = \mathsf{P}(U(\theta_j) = 0,\ e(\theta_j) = 1)$$

has a limit

$$\lim_{j\to\infty} h(j) = \frac{1}{\mathsf{E}\nu_1}.$$

Since $\{U(t), e(t)\}$ is an ergodic Markov chain and the processes $X(t)$ and $\{U(t), e(t)\}$ are independent, we have $(\mathsf{E}\nu_1)^{-1} = p_0^1 > 0$, and so $\mathsf{E}\nu_1 < \infty$. By Wald's identity, $\mathsf{E}\tau_1' = \mathsf{E}\tau_1\mathsf{E}\nu_1$, and we conclude that $\mathsf{E}\tau_1' < \infty$. □

Denote

$$x_n = X(T_n) - X(T_{n-1}), \qquad y_n = Y(T_n) - Y(T_{n-1}).$$

Since $\{T_n\}_{n=1}^{\infty}$ is a sequence of regeneration points common for $X(t)$ and $Y(t)$, it follows that the sequence $\{x_n, y_n\}_{n=1}^{\infty}$ consists of independent identically distributed random vectors, and we have $\mathsf{E}x_n < \infty$ and $\mathsf{E}y_n < \infty$ by Lemma 1.

Let $\mu(t)$ be the number of regenerations on $(0, t]$; i.e., $\mu(t) = \max\{k : T_k < t\}$.

We introduce the embedded process

$$W_n = W(T_n - 0)$$

and two auxiliary processes W_n^- and W_n^+ by setting

$$W_n^- = [W_{n-1}^- + x_n - y_n]^+, \qquad W_0^- = W(0),$$

$$W_n^+ = [W_{n-1}^+ - y_n]^+ + x_n, \qquad W_0^+ = W(0).$$

Then for each integer $n \geq 0$ we almost surely have

$$W_n^- \leq W_n \leq W_n^+. \tag{5}$$

It is well known (e.g., see [11]) that

$$W_n^- \xrightarrow[n\to\infty]{\text{a.s.}} \infty \qquad \text{if } \mathsf{E}(x_n - y_n) > 0,$$
$$W_n^- \xrightarrow[n\to\infty]{\mathsf{P}} \infty \qquad \text{if } \mathsf{E}(x_n - y_n) = 0 \quad \text{and} \quad \mathsf{P}\{x_n = y_n\} < 1.$$

Let us proceed to the case of $\mathsf{E}x_n < \mathsf{E}y_n$. It was shown in [6] that the process W_n^+ is stochastically bounded in this case. It follows from (5) that so is W_n. Note that $W(t)$ and W_n are regeneration processes with regeneration times $\widetilde{T}_n$ given by

$$\widetilde{T}_n = \min\{T_k > \widetilde{T}_{n-1} : W(T_k) = W_k = 0\}, \qquad \widetilde{T}_0 = 0.$$

By Theorem 1 in [12], the stochastic boundedness of $W(t)$ implies its ergodicity under the following conditions.

Condition 3. $\mathsf{P}(W_{n+1} = 0 \mid W_n = 0) > 0$.

Condition 4. For each $x < \infty$, $x \in \mathbb{R}$, there exists positive integers $m(x)$ and $\delta(x)$ such that

$$\mathsf{P}(W_{n+m(x)} = 0 \mid W_n = y) \geq \delta(x)$$

for all $y \leq x$.

Let us prove that these conditions are satisfied in our case.

Lemma 2. *It follows from Conditions* 1 *and* 2 *that Conditions* 3 *and* 4 *are satisfied.*

Proof. Assume for now that the first term in (1) is positive, $\mathsf{P}\{\xi_1 = 0, \tau_1 > 0\} > 0$. Then there exist $h_1 > 0$, $h_2 > 0$, and $\delta_1 > 0$ such that

$$\mathsf{P}\{\xi_1 = 0,\ h_1 < \tau_1 < h_2\} > \delta_1. \tag{6}$$

Since $\{e(t), U(t)\}$ is a Markov chain, it follows that the time spent in the state $\{1, 0\}$ is exponentially distributed with parameter $\gamma_0 = \sum_{j\neq 0} q_{0j}$. In view of (6), we find that

$$\mathsf{P}\{W_{n+1} = 0 \mid W_n = 0\} > e^{-\gamma_0 h_2}\delta_1. \tag{7}$$

For a given $x > 0$, take $m(x) = [x/h_1] + 1$. Then Condition 4 holds with $\delta(x) = e^{-\gamma_0 m(x) h_2}\delta_1^{m(x)}$.

Let the second term in (1) be positive. There exist $\widetilde{h}_1 > 0$, $\widetilde{h}_2 > 0$, and $\widetilde{\delta}_1 > 0$ such that

$$\mathsf{P}\{\xi_1 = 1, \eta_1 + t_1 + \widetilde{h}_1 < \tau_1 < \widetilde{h}_2\} > \widetilde{\delta}_1.$$

Hence (7) obviously holds with h_2 replaced by $\widetilde{h}_2$ and δ_1 replaced by $\widetilde{\delta}_1$. To prove that condition 4 is satisfied, it suffices to take $\widetilde{m}(x) = [x/\widetilde{h}_1] + 1$. Then

$$\mathsf{P}\{W_{n+\widetilde{m}(x)} = 0 \mid W_n = y\} \geq (e^{-\gamma_0 \widetilde{h}_2}\widetilde{\delta}_1)^{\widetilde{m}(x)} > 0$$

for $y \leq x$. □

It remains to express the system service factor $\rho = \mathsf{E}x_1/\mathsf{E}y_1$ via the original parameters.

Lemma 3. *The following equalities hold*:

$$\frac{\mathsf{E}x_1}{\mathsf{E}y_1} = \frac{\lambda b}{\pi} = \rho.$$

Proof. We have

$$\lim_{t\to\infty}\frac{X(t)}{t}=\lim_{t\to\infty}\frac{\mu(t)}{t}\frac{1}{\mu(t)}\sum_{n=1}^{\mu(t)}x_n.$$

Next,

$$\frac{\mu(t)}{t}\xrightarrow[t\to\infty]{\text{a.s.}}\frac{1}{\mathsf{E}\tau_1'}$$

by reconstruction theory and

$$\frac{1}{\mu(t)}\sum_{n=1}^{\mu(t)}x_n\xrightarrow[t\to\infty]{\text{a.s.}}\mathsf{E}x_1$$

by the law of large numbers. Thus,

$$\frac{X(t)}{t}\xrightarrow[t\to\infty]{\text{a.s.}}\frac{\mathsf{E}x_1}{\mathsf{E}\tau_1'}.$$

Likewise, we find that

$$\frac{Y(t)}{t}\xrightarrow[t\to\infty]{\text{a.s.}}\frac{\mathsf{E}y_1}{\mathsf{E}\tau_1'}.$$

Next, since

$$\frac{X(t)}{t}=\frac{1}{t}\sum_{n=1}^{A(t)}\eta_n\xrightarrow[t\to\infty]{\text{a.s.}}\lambda b,$$

it follows that $\mathsf{E}x_1=\lambda b\mathsf{E}\tau_1'$. For the process

$$Y(t)=\int_0^t 1(e(s)=1)\,ds,$$

we have

$$\lim_{t\to\infty}t^{-1}\mathsf{E}Y(t)=\lim_{t\to\infty}t^{-1}\int_0^t\mathsf{P}(e(s)=1)\,ds=\pi.$$

Thus, $\mathsf{E}y_1=\pi\mathsf{E}\tau_1'$, and consequently

$$\frac{\mathsf{E}x_1}{\mathsf{E}y_1}=\frac{\lambda b}{\pi},$$

where π is defined in (3). □

The proof of the theorem is complete. □

4. EXAMPLES

Example 1. Consider the above-mentioned prioritized system. There are two types of requests arriving in a single server queuing system, requests of the second type having absolute priority over those of the first. The input flows $A_i(t)$, $i=1,2$, are independent, $A_1(t)$ is a regenerative flow with arrival rate λ_1, and $A_2(t)$ is a Poisson flow with parameter λ_2. The service times for type 1 requests are independent identically distributed random variables with distribution function $B(x)$ and expectation b, and the service times for type 2 requests are distributed exponentially with parameter μ_2. Type 1 requests are only serviced when there are no type 2 requests in the system, and the interrupted service of a type 1 request is continued as soon as all type 2 requests leave the system.

Let $U(t)$ be the number of type 2 requests at time t. This is a Markov chain, and if $\rho_2 = \lambda_2/\mu_2 < 1$, then it has a steady-state distribution; moreover,

$$R_j = \lim_{t\to\infty} \mathsf{P}(U(t) = j) = (1-\rho_2)\rho_2^j.$$

The server interrupts servicing type 1 requests when type 2 requests arrive in the system; i.e., $\alpha_0 = 0$, $\beta_0 = 1$, and $\alpha_i = 1$ and $\beta_i = 0$ for $i \geq 1$, so that $\pi = R_0$ (see (3)) and the service factor has the form $\rho_1 = \lambda_1 b/(1-\rho_2)$. If $A_1(t)$ is a Poisson process, then we obtain the well-known ergodicity conditions (e.g., see [13]).

Example 2. Consider a $\mathrm{Reg}/G/1/\infty$ system operating in a random medium $U(t)$, where $U(t)$ is the number of requests in an $M/M/1/\infty$ system with Poisson input flow of intensity λ_1 and with exponentially distributed service time with parameter μ_1. It is well known [6] that $U(t)$ is ergodic if $\lambda_1 < \mu_1$. The infinitesimal matrix for $U(t)$ has the form

$$q_{ii+1} = \lambda_1, \quad i = 0, 1, \dots, \qquad q_{ii-1} = \mu_1, \quad i = 1, 2, \dots, \qquad q_{ij} = 0, \quad |i-j| > 1. \tag{8}$$

As $U(t)$ switches into an arbitrary state, the server breaks down with probability α if it is up and resumes operation with probability β if it is down. That is, $\alpha_i = \alpha$ and $\beta_i = \beta$ for $i \in \{0, 1, \dots\}$. We point out that the up- and downtime intervals are not independent in this system.

We introduce the notation

$$P_j = \lim_{t\to\infty} \mathsf{P}(U(t) = j,\, e(t) = 1), \qquad P(z) = \sum_{j=1}^{\infty} z^j P_j,$$

$$Q_j = \lim_{t\to\infty} \mathsf{P}(U(t) = j,\, e(t) = 0), \qquad Q(z) = \sum_{j=1}^{\infty} z^j Q_j,$$

$$R_j = P_j + Q_j, \qquad R(z) = \sum_{j=1}^{\infty} z^j R_j, \qquad \rho_1 = \frac{\lambda_1}{\mu_1}.$$

Now, to find the ergodicity condition for this system, we should find

$$\pi = \lim_{t\to\infty} \mathsf{P}(e(t) = 1) = \sum_{j=0}^{\infty} P_j.$$

To this end, we use the fact that

$$R_j = \lim_{t\to\infty} \mathsf{P}(U(t) = j) = (1-\rho_1)\rho_1^j$$

and write out the following system of equations using (8):

$$\begin{aligned} \lambda_1 P_0 &= (1-\alpha)\mu_1 P_1 + \beta\mu_1 Q_1, \\ (\lambda_1+\mu_1)P_j &= (1-\alpha)\lambda_1 P_{j-1} + (1-\alpha)\mu_1 P_{j+1} + \beta\lambda_1 Q_{j-1} + \beta\mu_1 Q_{j+1}, \qquad j = 1, 2, \dots. \end{aligned}$$

By setting $c = 1-\alpha-\beta$, we obviously obtain the relation

$$P(z)(-c\rho_1 z^2 + (1+\rho_1)z - c) = \frac{1-\rho_1}{1-\rho_1 z}\beta(1+\rho_1 z^2) + P_0(z-c) - \beta(1-\rho_1) \tag{9}$$

for the generating function $P(z)$. Hence

$$P(z) = \frac{(1-\rho_1)\beta(1+\rho_1 z^2) + (1-\rho_1 z)(P_0(z-c) - \beta(1-\rho_1))}{(-c\rho_1 z^2 + (1+\rho_1)z - c)(1-\rho_1 z)}.$$

To express P_0, we find the roots

$$z_{1,2} = \frac{(1+\rho_1) \pm \sqrt{(1+\rho_1)^2 - 4c^2\rho_1}}{2c\rho_1}$$

of the equation

$$g(z) = z^2 - \frac{1+\rho_1}{\rho_1 c}\, z + \frac{1}{\rho_1} = 0.$$

Of these two roots, we take z_2, because $z_2 \in (-1,1)$ and hence $P(z)$ is analytic at that point. In view of this, we use (9) and obtain

$$P_0 = \frac{\beta\rho_1(1-\rho_1)z_2(1+z_2)}{(c-z_2)(1-z_2\rho_1)}, \qquad P(1) = \frac{P_0}{1+\rho_1} + \frac{2\beta\rho_1}{(1-c)(1+\rho_1)}.$$

Since

$$P(1) = \sum_{j=0}^{\infty} P_j,$$

it follows that the service factor of the system is

$$\rho = \frac{\lambda b(\alpha+\beta)(1+\rho_1)}{(\alpha+\beta)P_0 + 2\beta\rho_1},$$

where λ is the arrival rate and b is the expectation of the service time.

Example 3. In this example, one cannot construct a Markov process $U(t)$ determining whether the server is up. However, one can find a sufficient ergodicity condition in an auxiliary model that operates in a Markov random medium and majorizes the original system. This condition is close to being necessary in the sense that if the system is not ergodic, then the condition is violated.

The model consists of two systems S_1 and S_2 arranged in a series (see the figure).

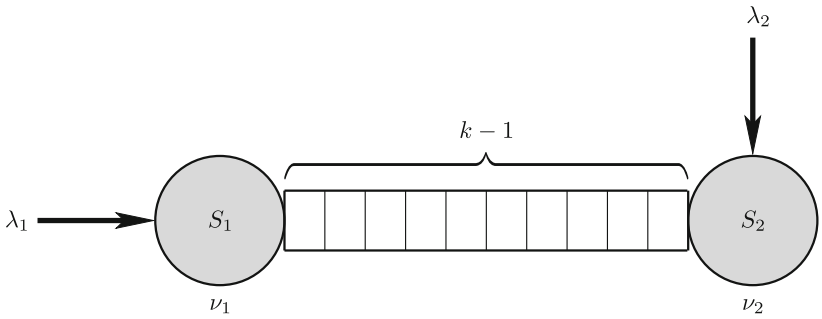

Figure.

Here S_1 is a single server queue with Poisson input flow with parameter λ_1. The service time is exponentially distributed with parameter ν_1. (Thus, S_1 is an $M/M/1/\infty$ system).

The system S_2 is a single server queue with two input flows. The first flow is formed by the requests that arrive from the system S_1. The second flow is Poisson with parameter λ_2. We assume that a request from the second flow is accepted by S_2 if the server is free and denied service otherwise. The service time in the system S_2 is exponentially distributed with parameter ν_2.

There is a buffer with $k-1$ cells, $k<\infty$, between S_1 and S_2. Accordingly, S_1 stops servicing requests as soon as the buffer is full, i.e., as soon as there are k requests in the system S_2.

Let $q_i(t)$ be the number of requests in S_i at time t, $i=1,2$. By virtue of our assumptions, $Q(t) = (q_1(t), q_2(t))$ is an irreducible Markov chain; since $q_2(t) \le k < \infty$, it follows that it is ergodic if and only if the process $q_1(t)$ is stochastically bounded [12].

Consider an auxiliary system $(\widetilde{S}_1, \widetilde{S}_2)$ assuming that there are backup requests in $\widetilde{S}_1$ which arrive in service as soon as the server does not have to service the main request arriving at the rate λ_1. If a main request arrives when the server is busy with a backup request, the server switches to the main request

immediately. Since we assume that there are as many backup requests as desired, we see that the input flow of $\widetilde{S}_2$ is Poisson with parameter ν_2. Of course, there will be delays if the buffer is full.

Let $\widetilde{q}_1(t)$ be the number of requests in $\widetilde{S}_1$ at time t, and let $\widetilde{q}_2(t)$ be the process $q_2(t)$ for $\widetilde{S}_2$. The $\widetilde{q}_2(t)$ is a Markov chain, and for the same initial conditions one has the stochastic inequality

$$q_1(t) \le \widetilde{q}_1(t) \qquad \text{as} \quad t \ge 0. \tag{10}$$

We can treat $\widetilde{q}_2(t)$ as a random medium for $\widetilde{S}_1$. Then

$$\alpha_0 = \cdots = \alpha_{k-1} = 0, \quad \alpha_k = 1, \qquad \beta_0 = \cdots = \beta_{k-1} = 1, \quad \beta_k = 0.$$

We conclude that the server in $\widetilde{S}_1$ is idle if $\widetilde{q}_2(t) = k$. By Theorem 1, the service ratio for $\widetilde{S}_1$ is given by

$$\widetilde{\rho} = \frac{\lambda_1}{\nu_1(1-\widetilde{R}_k)}, \qquad \text{where} \quad \widetilde{R}_k = \lim_{t\to\infty} \mathsf{P}(\widetilde{q}_2(t) = k).$$

For the steady-state distribution $\widetilde{R}_j = \lim_{t\to\infty} \mathsf{P}(\widetilde{q}_2(t) = j)$, we have the system of equations

$$\begin{aligned}
(\nu_1 + \lambda_2)\widetilde{R}_0 &= \nu_2\widetilde{R}_1,\\
(\nu_1 + \nu_2)\widetilde{R}_1 &= (\nu_1 + \lambda_2)\widetilde{R}_0 + \nu_2\widetilde{R}_2,\\
(\nu_1 + \nu_2)\widetilde{R}_j &= \nu_1\widetilde{R}_{j-1} + \nu_2\widetilde{R}_{j+1}, \qquad 1 < j < k,\\
\nu_2\widetilde{R}_k &= \nu_1\widetilde{R}_{k-1}.
\end{aligned}$$

Hence we find that

$$\widetilde{R}_k = \frac{\nu_1^{k-1}(\nu_1 + \lambda_2)(\nu_2 - \nu_1)}{\nu_2^k(\nu_2 - \nu_1) + (\nu_1 + \lambda_2)(\nu_2^k - \nu_1^k)}.$$

Thus, the service ratio is

$$\widetilde{\rho} = \frac{\lambda_1}{\nu_1\nu_2} \cdot \frac{\nu_2^{k+1} - \nu_1^{k+1} + \lambda_2(\nu_2^k - \nu_1^k)}{\nu_2^k - \nu_1^k + \lambda_2(\nu_2^{k-1} - \nu_1^{k-1})}.$$

If $\nu_1 = \nu_2$, the $\widetilde{R}_k = (\nu_1 + \lambda_2)/(\nu_1(k+1) + k\lambda_2)$ and the service ratio is

$$\widetilde{\rho} = \frac{\lambda_1}{\nu_1} \cdot \frac{\nu_1(k+1) + k\lambda_2}{k\nu_1 + \lambda_2(k-1)}.$$

Assume that $\widetilde{\rho} < 1$; i.e., the process $\widetilde{q}_1(t)$ is ergodic and hence stochastically bounded. It follows from (10) that $q_1(t)$ is stochastically bounded and hence ergodic.

Assume that $Q(t)$ is not ergodic. Then $q_1(t)$ is stochastically unbounded, and so is $\widetilde{q}_1(t)$ by (10). It follows that $\widetilde{\rho} \ge 1$.

Corollary 1. *If $\widetilde{\rho} < 1$, then $Q(t)$ is ergodic. If $Q(t)$ is not ergodic, then $\widetilde{\rho} \ge 1$.*

ACKNOWLEDGMENTS

The author wishes to express gratitude to Professor L. G. Afanas'eva for her lasting interest in the present research as well as for valuable remarks and advice, which served as a strong encouragement for writing the paper.

This work was supported by the Russian Foundation for Basic Research (grant no. 13-01-00653 A).

REFERENCES

1. H. White and L. S. Christie, "Queuing with preemptive priorities or with breakdown," Operations Res. **6** (1), 79–95 (1958).
2. D. P. Gaver, Jr., "A waiting line with interrupted service including priority," J. Roy. Statist. Soc. Ser. B **24**, 73–90 (1962).
3. J. Keilson, "Queues subject to service interruptions," Ann. Math. Statist. **33** (4), 1314–1322 (1962).
4. T. Kernane, *A Single Server Retrial Queue with Different Types of Server Interruptions*, E-print (2009).
5. A. Krishnamoorthy, P. K. Pramod, and T. G. Deepak, "On a queue with interruptions and repeat or resumption of service," Nonlinear Anal., Theory Methods Appl., Ser. A, Theory Methods **71** (12), e-Suppl., e1673–e1683 (2009).
6. L. G. Afanas'eva, "Queues with cyclic control processes," Kibernetika i Sistemnyi Analiz **41** (1), 54–68 (2005).
7. L. G. Afanasyeva and E. E. Bashtova, "Coupling method for asymptotic analysis of queues with regenerative input and unreliable server," Queueing Syst. **76** (2), 125–147 (2014).
8. R. A. Howard, "Research in semi-Markovian decision structures," J. Oper. Res. Soc. Japan **6** (4), 163–199 (1964).
9. L. G. Afanasyeva, E. E. Bashtova, and E. V. Bulinskaya, "Limit theorems for semi-Markov queues and their applications," Comm. Statist. Simulation Comput. **41** (6), 688–709 (2012).
10. L. G. Afanas'eva and E. V. Bulinskaya, *Stochastic Processes in the Theory of Queues and Inventory Control* (Izd. Moskov. Univ., Moscow, 1980) [in Russian].
11. A. A. Borovkov, *Stochastic Processes in Queueing Theory* (Nauka, Moscow, 1972) [in Russian].
12. L. G. Afanas'eva and A. V. Tkachenko, "Multichannel queueing systems with regenerative input flow," Teor. Veroyatnost. i Primenen. [Theory Probab. Appl.] **58** (2), 210–234 (2013) [Theory Probab. Appl. **58** (2), 174–192 (2013)].
13. T. L. Saaty, *Elements of Queueing Theory. With Applications* (McGraw-Hill, New York, 1961; Sovetskoe Radio, Moscow, 1971).