Problems of Information Transmission, Vol. 41, No. 3, 2005, pp. 254–279. Translated from Problemy Peredachi Informatsii, No. 3, 2005, pp. 76–104.

COMMUNICATION NETWORK THEORY

On the Stability of a Queueing System with Uncountably Branching Fluid Limits

A. P. Kovalevskii⋆[1], **V. A. Topchii**⋆⋆[2], **and S. G. Foss**⋆⋆⋆[3]

⋆*Novosibirsk State Technical University*
pandorra@online.nsk.su
⋆⋆*Omsk Branch of S.L. Sobolev Institute of Mathematics*
topchij@iitam.omsk.net.ru
⋆⋆⋆*S.L. Sobolev Institute of Mathematics, Novosibirsk*
Heriot-Watt University, Edinburgh, UK
foss@math.nsc.ru, foss@ma.hw.ac.uk

Received August 26, 2004; in final form, February 18, 2005

Abstract—It is known that in many queueing systems fluid limits are deterministic functions. Models and conditions which lead to random fluid limits have not received much attention. This paper is devoted to a study of a queueing network whose fluid limits admit a random and uncountable branching at certain points. Stability conditions for this model are investigated by the use of recent results from the theory of branching processes.

1. INTRODUCTION. THE FLUID APPROXIMATION APPROACH

Let $\{X(t),\ t \geq 0\}$ be a Markov process which describes the dynamics of a stochastic queueing network and possesses the strong Markov property, and let $|\cdot|$ be a certain nonnegative (test) function satisfying a number of "good" properties (say, a semi-norm). We provide here a short informal description of the *fluid approximation approach* for the study of stability (ergodicity), which was proposed originally in [1–3] and developed further by many authors (see, e.g., [4–9] and references therein).

Consider initial states x such that $|x| > 0$, and introduce a family of processes

$$X^x := \left\{X^x(t) = \frac{X(|x|t)}{|x|},\ t \geq 0\right\}, \quad X^x(0) = x.$$

Hereafter, the symbols := and =: are used for either the introduction of new notation or determination of values of the function standing adjacent to the colon.

Under certain intrinsic assumptions, for any fixed $T > 0$, the family of processes $\{X^x,\ |x| \geq 1\}$ is weakly compact in the space of cadlag functions $\mathcal{D}[0,T]$ with the Skorokhod topology. The latter means that any subsequence $x = x_n$, $|x_n| \to \infty$, contains a convergent subsubsequence such that the corresponding processes X^x converge in distribution in the Skorokhod metric to some

[1] Supported in part by the Russian Foundation for Basic Research, project no. 02-01-00358, and Grant of the President of the Russian Federation for Leading Scientific Schools, no. 2139.2003.1.
[2] Supported in part by the Russian Foundation for Basic Research, project no. 03-01-00045, Grant of the President of the Russian Federation for Leading Scientific Schools, no. 2139.2003.1, and Program 1.1 of the Branch of Mathematical Sciences of the Russian Academy of Sciences "Modern Problems of Theoretical Mathematics in the Institute of Mathematics of the Siberian Branch of the RAS."
[3] Supported in part by the Russian Foundation for Basic Research, project no. 02-01-00358, and EPSRC, grant no. T27099/01.

limiting process (say, φ), which is called a *fluid limit*. Fluid limits possess a number of good properties. In particular, almost all trajectories of any fluid limit are Lipshitz (with the same Lipshitz constant) and self-similar (see, e.g., [3]); i.e., for any fluid limit φ and for any constant $u > 0$ such that $\mathbf{P}(|\varphi(u)| > 0) > 0$, a random process $\{\widetilde{\varphi}(t),\ t \ge 0\}$ with conditional distribution

$$\mathbf{P}\left(\widetilde{\varphi}(t) \in \cdot\right) = \mathbf{P}\left(\frac{\varphi(u+t)}{|\varphi(u)|} \in \cdot \;\Big|\; \varphi(u)\right) \tag{1}$$

is also a fluid limit (for the underlying process X) a.s. on the event $\{|\varphi(u)| > 0\}$.

The study of stability of queueing networks attracts attention of many researchers. By the *stability* of a queueing network, we mean the *positive recurrence* of the process $\{X(t), t \ge 0\}$ (with respect to the test function $|\cdot|$), i.e., the existence of positive constants C and ε such that

$$\sup_{|x| \le C} \mathbf{E}\{\nu_C \mid X(0) = x\} < \infty,$$

where $\nu_C = \inf\{t \ge \varepsilon : |X(t)| \le C\}$. A queueing network is *unstable* if the process $X(t)$ is transient, i.e., at least one of the random variables ν_C is defective.

For a long time, the following hypothesis was considered to be "plausible":

(**H**) If the "local" traffic intensity is less than 1 at each station and if all service disciplines are "conservative" (i.e., a server cannot be idle if its queue is nonempty), then the whole network is necessarily stable.

The first examples which show that this hypothesis may, in general, be incorrect, were only found in the early 1990s. In particular, the papers [10,11] provide examples of deterministic systems where the hypothesis (**H**) fails.

Apparently, the article [1] is the first paper which studies examples of *stochastic* networks. The authors provide an example of a 2-station-2-customer-class network and show that (a) if at each station customers are served in the order of arrival then, (**H**) is valid; (b) but if, instead, certain priority disciplines are used, then (**H**) fails (an example of a system where the natural traffic conditions are not sufficient for stability even under FCFS disciplines was given later in [12]). In [1], the following three-step scheme for the proof of stability was proposed:

(1) First, the authors considered fluid limits and derived a family of integral-differential equations (IDEs) for a.s. all trajectories of all fluid limits;

(2) Then they found sufficient conditions for all solutions of these IDEs to vanish in a finite time; i.e., $|\varphi(t)| = 0$ for all sufficiently large t;

(3) Finally, they showed that the same conditions are sufficient for the positive recurrence of the underlying Markov process X.

However, many arguments in the proof at step (3) repeated the corresponding arguments of step (2).

In the subsequent papers [2,3], the approach was extended to a relatively large family of queueing networks (the so-called open multiclass networks). By the use of the *generalized Foster criterion*, it was shown that step (3) in the above scheme is superfluous. In particular, the following sufficient condition for stability was established:

(**C1**) There exists a constant $T \in (0, \infty)$ such that, for any fluid limit φ,

$$|\varphi(t)| = 0 \quad \text{a.s., for all } t \ge T. \tag{2}$$

Also, it was shown that condition (2) is equivalent to the following (formally weaker) condition:

(**C2**) There exist $\varepsilon \in (0,1)$ and, for any fluid limit φ, a constant $T' = T'(\varphi)$ (depending on φ) such that

$$|\varphi(T')| \le \varepsilon \quad \text{a.s.} \tag{3}$$

It was also shown in [3] that, if all fluid limits are deterministic functions, then either of the conditions (**C1**) and (**C2**) is equivalent to the following:

(**C3**) For any fluid limit φ,

$$\inf_{t\ge 0} |\varphi(t)| < 1.$$

It is known that in many basic queueing models the fluid limit is unique (up to an initial value) and is a deterministic function (this is valid, for instance, for generalized Jackson networks and for a cyclic polling system; see, e.g., [4–9]). This result may be understood by the following heuristic arguments:

(a) The input processes (i.e., the external arrival process of customers, service processes, etc.) in these networks are usually renewal; thus, the sequences of accumulated sums satisfy the strong law of large numbers (SLLN), which then lead—under linear scaling of time and space—to deterministic functions;

(b) If the control algorithm in such a network is "relatively simple," then it is reasonable to expect the fluid limit—as a function of the input and control—to be deterministic too.

The following question seems to be intrinsic: what kind of randomness may be preserved under passage to the fluid dynamics? As far as we are aware, this problem has been considered in only a small number of publications (see, e.g., [13–16]), where it was shown by examples that a fluid limit may be stochastic due to the following two causes: (1) randomness on a finite time horizon which is determined by an initial value; (2) randomness which is based on probabilistic branching (with a finite number of branches). In more detail, the fluid limits considered in these papers had the following properties:

(a) In the state space, one could choose a finite number of specific points (say, $\{x_i\}_{i=1}^{k}$, where $|x_i| = 1$) and, with each such point x_i, associate a certain discrete probability distribution $\{p_{i,j}\}_{j=1}^{\ell_i}$ and a corresponding family of deterministic fluid limits $\{\varphi_{i,j}\}_{j=1}^{\ell_i}$, each of which starts from the state $\varphi_{i,j}(0) = x_i$;

(b) For any fluid limit φ, one could define a sequence of finite stopping times $0 \le \tau_1 < \tau_2 < \ldots$ such that, for any $s \ge 2$,

$$\tau_s := \inf\{t > \tau_{s-1} :\ \varphi(t) \in D\},$$

where $D := \{cx_i\ :\ c > 0;\ i = 1, \ldots, k\}$, and, if $\varphi(\tau_s) = cx_i$ for some $c > 0$ and x_i, then with probability $p_{i,j}$ the trajectory $\{\varphi(\tau_s + u),\ 0 \le u \le \tau_{s+1} - \tau_s\}$ coincides with the trajectory of the deterministic process $\{c\varphi_{i,j}(u/c),\ 0 \le u \le \tau_{s+1} - \tau_s\}$;

(c) In the initial time interval $(0, \tau_1)$, a fluid limit may be random due to a specific characteristic of an initial value.

For models with (essentially) stochastic fluid limits, the stability conditions of criterion (**C1**) are only sufficient and far from necessary. For such models, wider stability conditions were proposed in [16].

Theorem 1. *Assume that all fluid limits satisfy the strong Markov property. If, for every fluid limit φ, there exists a stopping time τ_φ such that*

(a) *The random variables $\{\tau_\varphi\}$ are uniformly integrable* (*in the class of all fluid limits*);

(b) *For some $\varepsilon \in (0,1)$ and for all φ,*

$$|\varphi(\tau_\varphi)| \le \varepsilon \quad a.s., \tag{4}$$

then the underlying Markov process X is positive recurrent.

Condition (b) *is equivalent to the following condition*:

(b′) $\mathbf{E}\,|\varphi(\tau'_\varphi)| < \varepsilon$ *for some* (*other*) *family of uniformly integrable stopping times τ'_φ.*

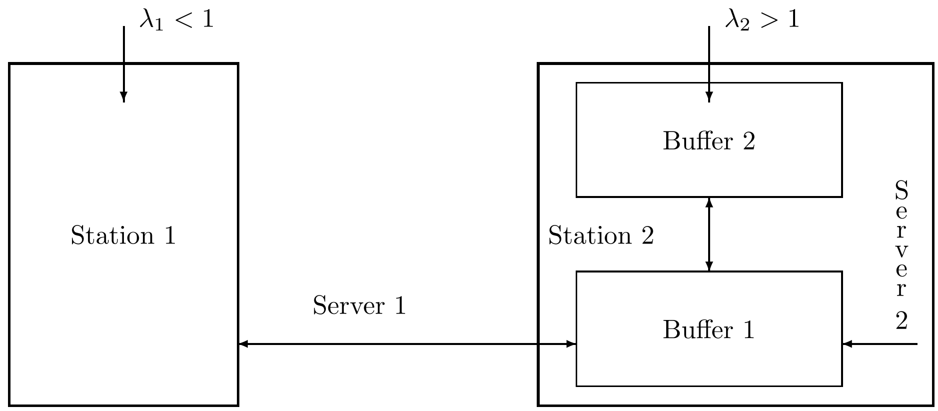


Fig. 1. Scheme of the queueing system. The buffers change places when one of them becomes idle; Server 1 assists Server 2 according to the algorithm described below.

Remark 1. Due to the strong Markov property, if there exists a family $\{\tau_\varphi\}$ satisfying (a) and (b) for *some* $\varepsilon > 0$, then such a family exists for *any* $\varepsilon \in (0,1)$.

In this paper, we propose a relatively simple and intrinsic model of a stochastic queueing network whose fluid limits stay random due to a slightly different cause: branching at a *single* point, but with a *uncountable* number of branches. With a certain complication, one can construct an example of a model with an uncountable branching at an uncountable number of points. The existence of uncountable branching of a limit process was proved in [14] for a random walk in $\mathbb{Z}_+^4$, and in [17] for a process of queue lengths in a loss system, but in both papers the probability measure on the family of branching trajectories was not properly defined. The impossibility of determining, in general, the limit (fluid) process leads to the absence of a complete classification of the asymptotic behavior of random walks in $\mathbb{Z}_+^4$. We note that the dynamics of our model may be described by a random walk on the boundary of $\mathbb{Z}_+^4$.

Our model is of interest because we are able to precisely describe its fluid limits with uncountable branching. For this, we use novel results for supercritical branching processes. Furthermore, we have met an intriguing new class of probability distributions, which we call "fractally exponential."

The structure of the paper is as follows. In Section 2, we give a description of the queueing system. In Section 3, we study an auxiliary nonergodic queueing system by methods from the theory of branching processes. In Section 4, we construct and study the fluid model and find stability and instability conditions for the original queueing system. Finally, the Appendix contains proofs of a number of auxiliary results.

2. DESCRIPTION AND PROPERTIES OF THE QUEUEING SYSTEM

The system consists of two stations and two servers. Station 1 receives an input stream of customers with rate $\lambda_1 < 1$, and Station 2 with rate $\lambda_2 > 1$; both service rates are equal to 1. Upon service completion, customers leave the system. The second server is always located at Station 2. Since its own service rate is less than that of the input, the first server assists it with service from time to time—along with its own service at the first station.

For the description of this assistance, we introduce three states for the queueing system.

The neutral state: each server works at its "own" station (the first server at the first station, the second server at the second station). The first server operates as the standard FCFS single server queue. The second station contains two buffers: the server serves the customers already accumulated in one of them, while new customers are accumulated in the other buffer. When the first buffer empties, the buffers change places: the "emptied" one starts to receive customers, while the server begins to serve customers from the "filled" buffer (if nothing has arrived at the "filled" buffer during the service of the other, the new service is assumed to be "fictitious": the server stays (without service) near the "filled" buffer and returns to the other buffer as soon as a new customer arrives there).

The state of readiness: the system operates in the same manner as in the neutral state, but the first server is ready for instantaneous transition to the second station.

The state of assistance: both servers operate together at the second station. They serve customers from one of the buffers, while new customers arrive at the other. During the service of the last customer from a buffer, one of the servers is free and just waits for the service completion. The buffers then change places. If there is only one customer in the "filled" buffer, it is served by the second server.

The system changes its state in the following cyclic order: neutral, readiness, assistance, then neutral again, etc. We describe now the transition algorithm from state to state.

The system stays in the neutral state until the moment when the first station becomes empty. The system then passes immediately to the state of readiness.

The system passes from the state of readiness to the state of assistance at the moment of change of places at the second station—given that there are at least two customers in the "filled" buffer (otherwise, the system stays in the state of readiness). If, at the time of transition to the state of assistance, there is a service in progress at the first station, it is interrupted and continues when the system passes to the neutral state again (and the first server returns to the first station).

The system returns to the neutral state when the second station becomes empty. If the first station is also empty, then the system immediately passes to the state of readiness.

Stochastic assumptions. We consider two variants of the stochastic assumptions.

(**SA1**) The inter-arrival intervals in the first input stream are i.i.d. with mean $1/\lambda_1$. The second input stream is Poisson with intensity λ_2 and does not depend on the first. For each server, its service times are i.i.d. with mean 1. Moreover, the typical service time σ of the second server has a finite moment $\mathbf{E}(\sigma \ln \sigma)$.

(**SA2**) The input streams are independent and Poisson with parameters λ_1 and λ_2. All service times are i.i.d. and have exponential distribution with parameter 1.

In Section 3, we obtain results under assumptions (**SA1**). Then in Section 4 we study the fluid model and its relationships with the underlying Markov process under the more particular assumptions (**SA2**) since the more general analysis follows the same scheme but requires many extra technicalities.

We are interested in (in)stability conditions for the queueing system. Throughout the paper, we assume that

$$\lambda_1 < 1 \qquad \text{and} \qquad \lambda_1 + \lambda_2 < 2 \tag{5}$$

since otherwise the system cannot be stable. Under condition (5), one can easily show that the system is always stable if $\lambda_2 \leq 1$. Therefore, we assume, along with (5), that

$$\lambda_2 > 1.$$

Note that, under the conditions $\lambda_1 < 1$ and $\lambda_2 < 2$, the inequality $\lambda_1 + \lambda_2 < 2$ is equivalent to

$$K := \frac{\lambda_1}{1-\lambda_1}\frac{\lambda_2 - 1}{2-\lambda_2} < 1.$$

Below, using fluid approximation, we find conditions for the stability and instability of our system. Simultaneously, we show that fluid limits admit a branching with uncountable number of branches.

Let $Q_1(t)$ be the number of customer at the first station at time t, and let $Q_{2,1}(t)$ and $Q_{2,2}(t)$ be the number of customers in the buffers at the second station. Let $Q_2(t) := Q_{2,1}(t) + Q_{2,2}(t)$. Assume that these processes are right-continuous.

For $n = 1, 2, \ldots$, consider the following sequence of (deterministic) initial states of the system:

$$Q_1^{(n)}(0) = X^{(n)}, \qquad Q_2^{(n)}(0) = 0, \quad \text{where} \quad X^{(n)} := X_1^{(n)} \to \infty.$$

We introduce some further notation. Let $R_0^{(n)} = 0$; let $T_1^{(n)}$ be the time of the first switch to the state of readiness, $U_1^{(n)} \geq T_1^{(n)}$ the time of the first switch to the state of assistance, $Y_1^{(n)}$ the total number of customers at the second station at this time, $Z_1^{(n)}$ the length of the first period in the state of assistance and $R_1^{(n)} := U_1^{(n)} + Z_1^{(n)}$ the time of the end of that period, and $X_2^{(n)}$ the number of customers at the first station at the time instant $R_1^{(n)}$. By induction, denote by $R_{k-1}^{(n)}$ the time of the $(k-1)$st switch to the neutral state and $X_k^{(n)}$ the number of customers at the first station at this moment. Then $R_{k-1}^{(n)} + T_k^{(n)}$ is the first subsequent time of switching to the state of readiness, $R_{k-1}^{(n)} + U_k^{(n)}$ the time of the kth switch to the state of assistance, $Y_k^{(n)}$ the total number of customers at the second station at time $R_{k-1}^{(n)} + U_k^{(n)}$, and $R_k^{(n)} := R_{k-1}^{(n)} + U_k^{(n)} + Z_k^{(n)}$. If $T_k^{(n)} > 0$, then let $V_k^{(n)} := U_k^{(n)}/T_k^{(n)}$.

Note that, if $X_1^{(n)} \to \infty$ as $n \to \infty$, then, for any fixed k, each of the characteristics introduced above (i.e., $X_k^{(n)}$, $T_k^{(n)}$, $U_k^{(n)}$, $Y_k^{(n)}$, and $Z_k^{(n)}$) also tends to infinity a.s. From the SLLN, for any fixed $k \in \mathbb{N}$ and as $n \to \infty$, the following convergence holds:

$$\frac{X_k^{(n)}}{T_k^{(n)}} \to 1 - \lambda_1, \qquad \frac{Y_k^{(n)}}{U_k^{(n)}} \to \lambda_2 - 1, \qquad \frac{Y_k^{(n)}}{Z_k^{(n)}} \to 2 - \lambda_2, \qquad \frac{X_{k+1}^{(n)}}{Z_k^{(n)}} \to \lambda_1. \tag{6}$$

The validity of each of these limiting relations is almost obvious. For instance, the first relation follows from the observations that at the time $R_{k-1}^{(n)}$ the first server switches to first station, there are $X_k^{(n)}$ customers there, the service rate is equal to 1, and new customers arrive with intensity $\lambda_1 < 1$. Therefore, the total queue length decreases with rate $1 - \lambda_1$.

Actually, Fig. 2 illustrates the limiting ($n = \infty$) case (under the intrinsic normalization). In the transient case, the trajectories are random curves which fluctuate in the neighborhood of corresponding straight lines. In particular, the horizontal segments at the zero level correspond to curves which fluctuate around zero.

For a complete description of fluid limits, we have to know possible limits of the ratios $V_k^{(n)} = U_k^{(n)}/T_k^{(n)}$ as $n \to \infty$. As we show at the end of Section 3, under certain rate of increase of $T_1^{(n)}$, these ratios converge to random variables. In particular, the number V_1 in Fig. 2 is random.

We start with a study of the asymptotic behavior of an auxiliary single-server queue.

3. AUXILIARY NON-ERGODIC QUEUEING SYSTEM AND BRANCHING PROCESSES

Throughout this section, we make assumptions (**SA1**); also, some particular results are formulated under assumptions (**SA2**).

The second station empties at time instants $R_k^{(n)}$. During the time intervals $(R_{k-1}^{(n)}, R_{k-1}^{(n)} + U_k^{(n)})$, it behaves as an ordinary $M/G/1$ queue which is overloaded: the input rate λ_2 is higher than the service rate 1. Consider the first of these time intervals. For simplicity of notation, we omit the index n.

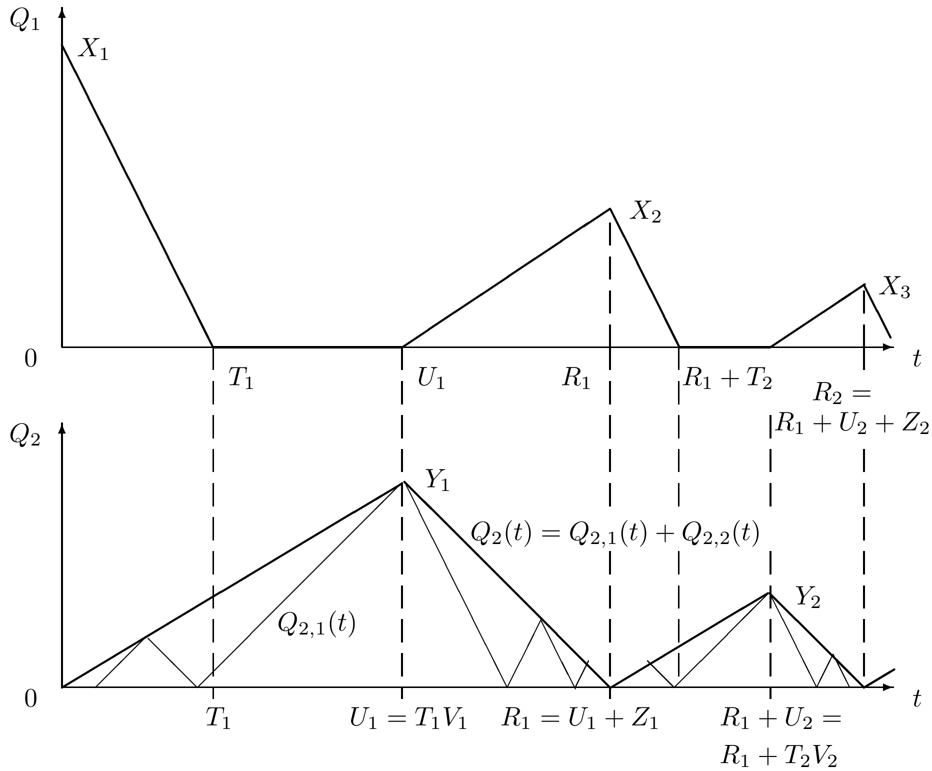


Fig. 2. Typical behavior of trajectories of the queue length processes at the first (top graph) and second (bottom graph) stations when $X_1 := X_1^{(n)} \gg 1$ (for all other characteristics, the upper index is omitted as well).

We call a *session* a period of service of a single buffer (or an idle time if the service is fictitious; in this case, the session is *empty*). For the ith session, let S_i be its length and ν_i the number of arrivals (at the second station). Clearly, the $(i+1)$st session contains the services of customers that arrive during the ith session. In particular, the $(i+1)$st session is empty if $\nu_i = 0$. Note that, for any i, $\nu_i = Q_2(S_0 + \ldots + S_i)$.

Let τ_1 be the arrival epoch of the first customer (at the second station) and τ_i, $i \ge 2$, the following inter-arrival times (these random variables are exponentially distributed with parameter λ_2). Further, let σ_ℓ be the ℓth service time at the second station. We assume that $\{\sigma_\ell\}$ are i.i.d. with mean 1 and that $\{\sigma_\ell\}$ and $\{\tau_\ell\}$ are independent. Let $\mu_i = \sum\limits_0^i \nu_j$, and let τ'_ℓ be the duration of the ℓth empty session. Then

(a) If $\nu_i > 0$, we have $S_{i+1} = \sum\limits_{j=\mu_{i-1}+1}^{\mu_i} \sigma_j$;

(b) The random variables τ'_ℓ have also an exponential distribution with parameter λ_2; if $\nu_{i-1} = 0$, we have $S_i = \tau'_i$ for the corresponding ℓ and, further, $\nu_1 = 1$ and $S_{i+1} = \sigma_{\mu_i}$;

(c) In particular, $\nu_0 = 1$ and $S_0 = \tau_1 = \tau'_1$.

One can easily find the distribution of the random variable ν_1: for $k = 0, 1, \ldots$, we have

$$\mathbf{P}\{Q_2(S_0+S_1) = k\} = \mathbf{E}\,\mathbf{P}\{Q_2(S_0+S_1) = k \,|\, \sigma_1\} = \int\limits_0^\infty \frac{(\lambda_2 t)^k}{k!} e^{-\lambda_2 t}\, dF_\sigma(t),$$

where $F_\sigma(t)$ is the distribution function of σ_1.

If, in particular, the σ_ℓ are exponentially distributed, then

$$\mathbf{P}\{Q_2(S_0+S_1)=k\}=\int\limits_0^\infty \frac{(\lambda_2 t)^k}{k!}e^{-\lambda_2 t}e^{-t}\,dt=\frac{\lambda_2^k}{(\lambda_2+1)^{k+1}}.$$

Clearly, the generating function of ν_1 may be represented as

$$f(z):=\sum_{k=0}^\infty z^k\int\limits_0^\infty \frac{(\lambda_2 t)^k}{k!}e^{-\lambda_2 t}\,dF_\sigma(t)=\int\limits_0^\infty e^{(z-1)\lambda_2 t}\,dF_\sigma(t)=\varphi_\sigma((1-z)\lambda_2)$$

where $\varphi_\sigma(\lambda)$ is the Laplace transform of σ_1.

The latter calculations imply the following statement.

Lemma 1. *The random variables $\nu_0,\nu_1,\ldots$ form a supercritical branching process with immigration at zero, where $\nu_0=1$;*

The random variable ν_1 has a discrete distribution with generating function $f(z)=\varphi_\sigma((1-z)\lambda_2)$ (in the case where σ_1 is exponentially distributed, its generating function is equal to $f(z)=(1+\lambda_2(1-z))^{-1}$) with mean $\mathbf{E}\,\nu_1=\lambda_2>1$;

If $\nu_{k-1}>0$, then ν_k admits the representation $\nu_k=\nu_k^{(1)}+\ldots+\nu_k^{(\nu_{k-1})}$, where $\nu_i^{(j)}$, $i,j=1,2,\ldots$, are i.i.d. random variables with generating function $f(z)$;

If $\nu_{k-1}=0$, then $\nu_k=1$.

Indeed, the random number of arrivals during any service time has the same distribution as ν_1, and, from the properties of Poisson processes, these random variables are independent.

Define the Galton–Watson process $N_0,N_1,\ldots$ which differs from the original process $\nu_0,\nu_1,\ldots$ in the absence of immigration at zero only: $N_0=1$; $N_1=N_1^{(1)}$ has a discrete distribution with generating function $f(z)=\varphi_\sigma((1-z)\lambda_2)$ and $\mathbf{E}\,N_1=\lambda_2>1$; if $N_{k-1}>0$, then $N_k=N_k^{(1)}+\ldots+N_k^{(N_{k-1})}$, where the random variables $N_i^{(j)}$, $i,j=1,2,\ldots$, are independent and distributed as N_1; if $N_{k-1}=0$, then $N_k=N_{k+1}=\ldots=0$.

The sequence $N_k/{\lambda_2}^k$ forms a nonnegative martingale, which converges a.s. to a limit W (see [18, ch. I, Section 8] or [19, ch. II, Section 1]). The distribution of W has an atom at zero. Its mass, q, is the probability of extinction of the process. It is also the unique solution of the equation $q=f(q)=\varphi_\sigma((1-q)\lambda_2)$ on the interval $[0,1)$ (if σ_1 has an exponential distribution, then $q=\lambda_2^{-1}$) if and only if the moment $\mathbf{E}\,N_1\ln N_1$ is finite. In this case, W has a continuous positive density on the positive half-line (see [19, ch. II, Sections 1 and 2; ch. III, Sections 5 and 6]). The moment generating function $\varphi(s)$ of W satisfies the relation $\varphi(\lambda_2 s)=f(\varphi(s))$ [19, ch. III, Section 5]. In addition, the explicit form of its density may be found analytically only if probabilities from the generating function $f(z)$ form a geometric progression [18, ch. I, Section 8]. In our case, the latter means that σ_1 is exponentially distributed and, if one lets $\alpha=(\lambda_2-1)/\lambda_2$, then $\mathbf{P}\{W\ge x\}=\alpha\exp\{-\alpha x\}$ for $x>0$.

Introduce a random variable W_0 with the distribution

$$\mathbf{P}\{W_0\ge x\}=\mathbf{P}\{W\ge x\,|\,W>0\}.$$

Denote by $m_1<m_2<\ldots<m_r$ the successive indices of generations k with $\nu_k=0$; i.e., $\nu_{m_1}=\nu_{m_2}=\ldots=\nu_{m_r}=0$ and $\nu_k\ne 0$ for $k\ne m_j$, $j=1,\ldots,r$. Then r, the number of zeros of the process, and $m_1,m_2,\ldots,m_r$ are proper random variables, and m_r is the last time instant when the process takes a zero value. Let $m_r=-1$ if $r=0$. Note that $m_r\ne 0$ with probability one.

Since the probability of extinction of the Galton–Watson process is q (see [18]) and the number r of degenerations of the branching process with immigration at zero coincides with the number of independent trials until the first success with a probability of success $1-q$, we have $\mathbf{P}\{r=k\} = q^k(1-q)$, $k=0,1,\ldots$. Hence, m_r is finite a.s.

Lemma 2. *As $k\to\infty$, the fraction ν_k/λ_2^k converges a.s. to a random variable ζ with distribution*

$$\mathbf{P}\{\zeta \ge t\} = \sum_{j=-1}^{\infty} \mathbf{P}\{W_0 \ge t\lambda_2^{j+1}\}\,\mathbf{P}\{m_r = j\}. \tag{7}$$

In particular, $\mathbf{P}\{\zeta = 0\} = 0$.

Proof. For $t\ge 0$, $j\ge -1$ ($j\ne 0$), and $k>j+1$,

$$\mathbf{P}\Big\{\frac{\nu_k}{\lambda_2^k} \ge t \,\Big|\, m_r = j\Big\} = \mathbf{P}\Big\{\frac{\nu_k}{\lambda_2^k} \ge t \,\Big|\, \nu_j = 0,\ \nu_{j+1}=1,\ \nu_{j+i}>0,\ i=2,3,\ldots\Big\}$$
$$= \mathbf{P}\left\{\frac{N_{k-j-1}}{\lambda_2^k} \ge t \,\Big|\, N_1>0,\ N_2>0,\ldots\right\} = \mathbf{P}\left\{\frac{N_{k-j-1}}{\lambda_2^{k-j-1}} \ge t\lambda_2^{j+1} \,\Big|\, N_1>0,\ N_2>0,\ldots\right\}.$$

Then, as $k\to\infty$,

$$\mathbf{P}\Big\{\frac{\nu_k}{\lambda_2^k} \ge t \,\Big|\, m_r = j\Big\} \to \mathbf{P}\left\{W \ge t\lambda_2^{j+1} \,\Big|\, W>0\right\},$$

$$\mathbf{P}\Big\{\frac{\nu_k}{\lambda_2^k} \ge t\Big\} = \sum_{j=-1}^{\infty} \mathbf{P}\Big\{\frac{\nu_k}{\lambda_2^k} \ge t \,\Big|\, m_r = j\Big\}\mathbf{P}\{m_r=j\} \to \sum_{j=-1}^{\infty} \mathbf{P}\left\{W \ge t\lambda_2^{j+1} \,\Big|\, W>0\right\}\mathbf{P}\{m_r=j\}$$

since m_r is a proper random variable.

Note that $\mathbf{E}\{\nu_k \,|\, \nu_{k-1}=c\} = c\lambda_2$ for $c>0$, and $\mathbf{E}\{\nu_k \,|\, \nu_{k-1}=0\} = 1$.

The sequence

$$V_k := \frac{\nu_k}{\lambda_2^k}$$

forms a submartingale. Indeed, $\mathbf{E}\{V_k \,|\, V_{k-1}=a\} = a$ for $a>0$, and $\mathbf{E}\{V_k \,|\, V_{k-1}=0\} = \lambda_2^{-k}$.

Clearly, $\nu_k \le \sum\limits_{i=1}^{k} N_i^{(i)}$, where the random variables $N_i^{(i)}$ are independent and distributed as N_i.

Then $\mathbf{E}\,V_k \le \sum\limits_{i=1}^{k} \mathbf{E}\,N_i^{(i)}\lambda_2^{-k} = \sum\limits_{i=1}^{k} \lambda_2^{-k+i} \le \lambda_2(\lambda_2-1)^{-1}$.

From Doob's theorem on the convergence of submartingales (see, e.g., [22, ch. 7, Section 4, Theorem 1]), the random variables V_k converge a.s. to a random variable ζ. △

Let $\Sigma^0_{(k)} := \nu_0+\nu_1+\ldots+\nu_k$ and $\Sigma_{(k)} := S_0+S_1+\ldots+S_k$.

Lemma 3. *As $n\to\infty$, we have $\Sigma^0_{(n)}\lambda_2^{-n} \xrightarrow{\text{a.s.}} \zeta/\alpha$ and $\Sigma_{(n+1)}\lambda_2^{-n} \xrightarrow{\text{a.s.}} \zeta/\alpha$, where $\alpha := 1-\lambda_2^{-1}$.*

Proof. From Lemma 2, we have

$$\frac{\Sigma^0_{(k)}}{\lambda^k} = \sum_{i=1}^{k} \frac{\nu_i}{\lambda_2^i\lambda_2^{k-i}} \xrightarrow{\text{a.s.}} \zeta\sum_{i=0}^{\infty}\lambda_2^{-i} = \zeta\frac{\lambda_2}{\lambda_2-1} = \frac{\zeta}{\alpha}.$$

Consider the ratio

$$\frac{\Sigma_{(k)}}{\Sigma^0_{(k-1)}} = \frac{\sum\limits_{j=1}^{\Sigma^0_{(k-1)}} \sigma_j + \sum\limits_{j=0}^{r(k-1)} \tau'_{m_j}}{\Sigma^0_{(k-1)}},$$

where the random variable $r(k)$, which is equal to the number of zeros in the sequence $\nu_1, \ldots, \nu_k$, is bounded by the geometrically distributed random variable r defined above (i.e., $r(k) \le r$ a.s.). From the first argument of the proof, $\Sigma^0_{(k)} \xrightarrow{\text{a.s.}} +\infty$. Therefore, from the SLLN, $\Sigma_{(k)}/\Sigma^0_{(k-1)} \xrightarrow{\text{a.s.}} \mathbf{E}\,\sigma_1 = 1$. △

We now apply these results to the limiting scheme which was introduced at the end of Section 2. Fix a number $H > 0$ and consider a sequence $X_1^{(n)}$ such that $T_1^{(n)}/\lambda_2^{n-1} \to H$.

Introduce the following notation:

$$\widetilde{U}_1^{(n)} := \min\left\{\Sigma_{(j)} : \Sigma_{(j)} \ge T_1^{(n)}\right\}, \qquad \eta^{(n)} := \min\left\{j : \Sigma_{(j)} \ge T_1^{(n)}\right\}.$$

Then $\widetilde{U}_1^{(n)} = \Sigma_{(\eta^{(n)})}$. The random variable

$$U_1^{(n)} = \min\left\{S_0 + \ldots + S_j : S_0 + \ldots + S_j \ge T_1^{(n)},\ Q_2(S_0 + \ldots + S_j) \ge 2\right\}$$

(which was introduced earlier in Section 2) can be rewritten as

$$U_1^{(n)} = \min\left\{\Sigma_{(j)} : \Sigma_{(j)} \ge T_1^{(n)},\ \nu_j \ge 2\right\}.$$

Theorem 2. *For any sequence* $T_1^{(n)}$ *such that* $T_1^{(n)}/\lambda_2^{n-1} \to H > 0$, *we have the convergence* $U_1^{(n)}/T_1^{(n)} \to V(H)$ *a.s. Here the random variable* $V(H)$ *takes values in the interval* $[1, \lambda_2]$ *and has the following distribution function*:

$$F_{V(H)}(t) = \sum_{k=-\infty}^{\infty} F_{W_0}(\alpha^2 \lambda_2^k H t) - F_{W_0}(\alpha^2 \lambda_2^k H), \quad t \in [1, \lambda_2]. \tag{8}$$

In particular, in the case of exponentially distributed service times,

$$F_{V(H)}(t) = \sum_{k=-\infty}^{\infty} \left(e^{-\alpha^2 \lambda_2^k H} - e^{-\alpha^2 \lambda_2^k H t}\right), \quad t \in [1, \lambda_2]. \tag{9}$$

Proof. From Lemma 3, as $n \to \infty$, the conditions $T_1^{(n)} \to \infty$ and $\Sigma_j \ge T_1^{(n)}$ imply that $j \to \infty$. Therefore, from Lemma 2, we have $\nu_j \xrightarrow{\text{a.s.}} \infty$, and the condition $\nu_j \ge 2$ in the definition of $U_1^{(n)}$ may be omitted, i.e., $\widetilde{U}_1^{(n)} - U_1^{(n)} \xrightarrow{\text{a.s.}} 0$. Thus, it suffices to show that $\widetilde{U}_1^{(n)}/T_1^{(n)} \to V(H)$ a.s. and to get relations (8) and (9) for $\widetilde{U}_1^{(n)}$.

It is sufficient to consider the case $1 \le H \le \lambda_2$ only. Indeed, for any fixed $k \in \mathbb{Z}$ and $1 \le H \le \lambda_2$ and as $n \to \infty$, for the sequence $\widetilde{U}_1^{(n)}$, the relative limits of the overshoots $V(H)$ and $V(H\lambda_2^k)$, $1 \le H \le \lambda_2$, over the corresponding levels $\lambda_2^n H$ and $\lambda_2^{n+k} H$ coincide. Then, due to the identity $V(H) \equiv V(\lambda_2 H)$, the statement of the theorem becomes valid for any $0 < H < \infty$. In particular, the first condition of Theorem 2 may be replaced by $T_1^{(n)}/\lambda_2^n \to H > 0$.

From Lemma 3, the sequence Σ_k satisfies the conditions of Theorems 1 and 2 of [21]; i.e., the convergence $\Sigma_n/m_n \to \zeta/(\alpha H)$ holds a.s., where $m_n := \lambda_2^n$ is a strictly increasing sequence and $m_{n+1}/m_n = \lambda_2 > 1$. Then, from [21, Theorem 1], we have

$$\eta^{(n)} - n \xrightarrow{\text{a.s.}} \left[-\frac{\ln \frac{\zeta}{\alpha H}}{\ln \lambda_2}\right] =: \eta_H.$$

From [21, Theorem 2], we have

$$\left(\eta^{(n)} - n,\ \widetilde{U}_1^{(n)} \lambda_2^{-n}\right) \xrightarrow{\text{a.s.}} \left(\eta_H, \frac{\zeta \lambda_2^{\eta_H}}{\alpha}\right).$$

Hence,

$$\mathbf{P}\{V(H) \geq t\} = \sum_{k=-\infty}^{\infty} \mathbf{P}\left\{\zeta\lambda_2^k \geq t\alpha H,\, \eta_H = k\right\} = \sum_{k=-\infty}^{\infty} \mathbf{P}\left\{t\alpha H\lambda_2^{-k} \leq \zeta < H\alpha\lambda_2^{-k+1}\right\}. \quad (10)$$

Thus, (7) and (10) imply that

$$\begin{aligned}\mathbf{P}\{V(H) \geq t\} &= \sum_{k=-\infty}^{\infty}\sum_{j=-1}^{\infty} \mathbf{P}\left\{tH\lambda_2^{j-k+1}\alpha \leq W_0 < H\lambda_2^{j-k+2}\alpha\right\}\mathbf{P}\{k_r = j\} \\ &= \sum_{s=-\infty}^{\infty} \mathbf{P}\left\{tH\lambda_2^{s}\alpha \leq W_0 < H\lambda_2^{s+1}\alpha\right\}\sum_{j=-1}^{\infty}\mathbf{P}\{k_r = j\} \\ &= \sum_{s=-\infty}^{\infty} \mathbf{P}\left\{tH\lambda_2^{s}\alpha \leq W_0 < H\lambda_2^{s+1}\alpha\right\}. \quad \triangle\end{aligned}$$

In particular, in the case of exponentially distributed service times, the random variable $V(H)$ has the following density:

$$f_{V(H)}(t) = \sum_{k=-\infty}^{\infty} \alpha^2\lambda_2^k H \exp\left\{-\alpha^2\lambda_2^k Ht\right\}, \quad t \in [1;\lambda_2]. \quad (11)$$

It is natural to call this distribution "fractally exponential."

Remark 2. As was already mentioned, the distribution of W_0 has a density which is strictly positive everywhere on the positive half-line. From formula (8), for any positive H, the random variable $V(H)$ has a density H which is bounded away from zero (uniformly over H). In equation (10), not only the probabilities but also the corresponding events coincide (see [21] for details). Therefore, the joint distributions of the process $\{V(H),\, 1 \leq H < \lambda_2\}$ are uniquely defined; thus, from the periodicity, the joint distributions of the family of random variables $\{V(H),\, 0 < H < \infty\}$ are determined as well. From (8) and the absolute continuity of the distribution of W_0, we get that the distribution of $V(H)$ depends continuously on the parameter $H > 0$. The latter implies that, as $n \to \infty$, if $H^{(n)} \to H > 0$, then $V(H^{(n)}) \to V(H)$ a.s., and if the sequence $T_1^{(n)}$ has the form $T_1^{(n)} = \lambda_2^{(n)}H^{(n)}$, then

$$\left|\frac{U_1^{(n)}}{T_1^{(n)}} - V\left(H^{(n)}\right)\right| \to 0 \quad \text{a.s.}$$

Moreover, from the first part of this remark, one can easily obtain (using a proof by contradiction) that the latter convergence holds for any sequence $H^{(n)}$ such that $\inf H^{(n)} > 0$ and $\sup H^{(n)} < \infty$ (i.e., it is not necessary to assume that the sequence converges to a limit H). Finally, we note that the family $\{V(H)\}$ satisfies the following *monotonicity* property: if $H_1 < H_2$, then $H_1V(H_1) \leq H_2V(H_2)$ a.s.

From Lemma 7 and the estimates (19), it follows that, under assumptions (**SA2**), although the distributions of $V(H)$ differ for different H, they are very close to each other (the distributions of the logarithms of these random variables have distributions that are very close to the uniform distribution on $[0, 1]$).

We define now fluid limits for the two-station queueing system and study their properties.

4. PROPERTIES OF THE FLUID MODEL

A *fluid model* (see, e.g., [2,3]) is a family of weak limits of the process in the uniform metric in any fixed time interval $[0;T]$ under the linear scaling by the initial value both in time and in space.

As will follow from the construction below, for our queueing system a stronger form of tightness holds: all *fluid limits* are a.s. limits of corresponding sequences of processes in the sense of the uniform convergence in $\mathcal{D}[0;T]$, for any fixed $T > 0$.

We assume here that assumptions (**SA2**) hold. Then the state of a system at time instant t is described by a right-continuous strictly Markov process $\big(Q^{(n)}(t); A^{(n)}(t)\big)$, where

$$Q^{(n)}(t) := \left(Q_1^{(n)}(t);\ Q_{2,1}^{(n)}(t);\ Q_{2,2}^{(n)}(t)\right),$$

with initial state $Q^{(n)}(0)$, and the discrete-state process $A^{(n)}(t) := \big(A_1^{(n)}(t); A_2^{(n)}(t)\big)$ characterizes the position of servers. Here $A_1 \in \{1;2\}$ is the number of the station where the first server is present, and $A_2 \in \{0;1;2\}$ is the number of the buffer under service; we put $A_2 = 0$ if the second station is empty (if this happens, the second server is inactive). If the whole system is empty (of customers), the first server stays at the first station. All states of the Markov process form a single class of intercommunicating states.

For the process $Q^{(n)}$, introduce the norm $|Q^{(n)}(t)| := Q_1^{(n)}(t) + Q_{2,1}^{(n)}(t) + Q_{2,2}^{(n)}(t)$.

Consider initial states of the form $Q^{(n)}(0) = (X_1^{(n)}; 0; 0)$, where $X_1^{(n)} = (1-\lambda_1)H^{(n)}\lambda_2^n$ and $\{H^{(n)}\}$ is a sequence of positive numbers which converges to $H > 0$ as $n \to \infty$.

For convenience, when passing to fluid limits, we normalize the queue length process not by the values $X_1^{(n)}$ but instead by numbers λ_2^n, proportional to them.

More precisely, we consider processes that are limits of the sequences

$$\frac{Q^{(n)}(\lambda_2^n t)}{\lambda_2^n}.$$

For any fixed $H > 0$ we consider a certain fluid limit $(q_1(t), q_{2,1}(t), q_{2,2}(t))$, $t \ge 0$, whose distributions depend on H only and (given H) do not depend on the sequence H_n. Recall that, by a *fluid model*, we mean the family of all limiting processes that depend on the parameter H.

From the general properties of fluid limits (see, e.g., [16, Theorem 1]), convergence of the processes

$$\frac{Q^{(n)}(\lambda_2^n t)}{\lambda_2^n}$$

to a limiting one is *uniform* on the interval $[0;T]$, for any $T > 0$.

We now fix a positive number $H > 0$ and describe the corresponding fluid limit. More precisely, we need to consider the two-dimensional process $(q_1(t), q_2(t))$, where $q_2(t) := q_{2,1}(t) + q_{2,2}(t)$.

We use the notation introduced in Section 2 and relations (6). We have

$$\frac{T_1^{(n)}}{\lambda_2^n} \to H =: t_1 \quad \text{a.s.}$$

Moreover, from the SLLN, for $t \in [0, t_1]$,

$$\frac{Q_1^{(n)}(\lambda_2^n t)}{\lambda_2^n} \to (1-\lambda_1)(H-t) =: q_1(t) \quad \text{a.s.}$$

and

$$\frac{Q_2^{(n)}(\lambda_2^n t)}{\lambda_2^n} \to (\lambda_2 - 1)t =: q_2(t) \quad \text{a.s.}$$

Further, from Theorem 2,

$$\frac{U_1^{(n)}}{\lambda_2^n} \to HV(H) =: u_1,$$

where $V(H)$ is a random variable with the distribution (9). For $t \in (t_1, u_1)$, we have

$$q_1(t) := 0 \qquad \text{and} \qquad q_2(t) := (\lambda_2 - 1)t$$

since the first server continues to work at the first station, where the traffic intensity is less than 1. In particular,

$$q_2(u_1) := y_1 := \lim \frac{Y_1^{(n)}}{\lambda_2^n} = (\lambda_2 - 1)u_1 \quad \text{a.s.},$$

where $Y_1^{(n)} := Q_2^{(n)}(U_1^{(n)})$.

Further,

$$\frac{R_1^{(n)}}{\lambda_2^n} \to r_1 := u_1 + z_1,$$

where

$$z_1 := \lim \frac{Z_1^{(n)}}{\lambda_2^n} = \frac{y_1}{2 - \lambda_2} \quad \text{a.s.}$$

For $t \in (u_1, r_1)$, we have

$$q_1(t) := \lambda_1(t - u_1)$$

and

$$q_2(t) := y_1 - (2 - \lambda_2)(t - u_1).$$

In particular,

$$q_1(t_1) := x_2 := \lim \frac{X_2^{(n)}}{\lambda_2^n} = \lambda_1 z_1 \qquad \text{and} \qquad q_2(r_1) := 0.$$

Finally,

$$t_2 := \lim \frac{T_2^{(n)}}{\lambda_2^n} = \frac{x_2}{1 - \lambda_1} = t_1 V(t_1) K \quad \text{a.s.}$$

We now use induction arguments for the construction of a fluid limit. For this, assume that $\{V_i(H),\, H > 0\}_{i \ge 1}$ is a sequence of i.i.d. copies of the family $\{V(H),\, H > 0\}$. Then, for any $i \ge 1$,

$$u_i := t_i V_i(t_i), \qquad y_i := (\lambda_2 - 1)y_i, \qquad z_i := \frac{y_i}{2 - \lambda_2}, \qquad x_{i+1} := \lambda_1 z_i, \qquad t_{i+1} := \frac{x_{i+1}}{1 - \lambda_1};$$

in particular,

$$t_{i+1} := t_i V_i(t_i) K.$$

Here $q_1(t)$ and $q_2(t)$ are a.s. continuous and piecewise linear functions.

Remark 3. For the study of positive recurrence, it is sufficient to consider initial values of the form $Q^{(n)}(0) = (X_1^{(n)}; 0; 0)$ only. Indeed, if $Q^{(n)}(0) \ne (X_1^{(n)}; 0; 0)$, then the first passage time $\eta^{(n)}$ to the subspace $(\cdot\,; 0; 0)$ satisfies the condition $\eta^{(n)}/R_1^{(n)} \le 1$ a.s. Assume that $\eta^{(n)}/T_1^{(n)} \to \eta$ a.s. Then

$$\eta \le \frac{u_1 + z_1}{t_1} \le \lambda_2 \left(1 + \frac{\lambda_2 - 1}{2 - \lambda_2}\right)$$

a.s. since $u_1/t_1 \le \lambda_2$ a.s. due to Theorem 2.

The trajectories of the limiting process are a.s. continuous piecewise linear functions, and the maximum and the minimum of the norm on a cycle are attained, respectively, at the points t_1 and u_1. Here both the maximum and minimum are strictly positive and finite. Therefore, there exist positive constants $C_1 < C_2$ such that $0 < C_1 \le |q(\eta)|/|q(0)| \le C_2 < \infty$ a.s.

Thus, the norm of the limiting process at the time of the first passage to the subspace $(\cdot\,;0;0)$ is bounded from above and below by positive constants, and the first passage time itself is bounded from above. Therefore, in the sequel, we consider only initial states $(\cdot\,;0;0)$.

Let $t_1 = H$. Let, for some $\varepsilon \in (0,1)$,

$$\mu := \mu_H(\varepsilon) := \min\{n \ge 2 :\ t_n \le \varepsilon H\}$$

and

$$\tau := \tau_H(\varepsilon) := r_{\mu-1} + t_\mu.$$

The conditions of Theorem 1 will be satisfied if we prove that the sequence of random variables $\{t_n\}$ decreases sufficiently rapidly, namely, in such a way that the family of random variables $\{\tau_H,\ 1 \le H < \lambda_2\}$ is uniformly integrable. From

$$t_n \le r_n - r_{n-1} \le t_n \frac{\lambda_2}{2-\lambda_2} \quad \text{a.s.},$$

we get

$$\varkappa := \varkappa_H(\varepsilon) := \sum_1^{\mu} t_i \le \tau \le \frac{\lambda_2}{2-\lambda_2}\varkappa,$$

and the underlying queueing system is stable if, for some $\varepsilon \in (0,1)$, the family of random variables

$$\{\varkappa_H(\varepsilon), H \in [1,\lambda_2)\} \quad \text{is uniformly integrable.} \tag{12}$$

It is easy to show that if condition (12) holds for *some* $\varepsilon \in (0,1)$, then it holds for *all* $\varepsilon \in (0,1)$.

Let $t_1 = H \in [1,\lambda)$ and $\theta_H := \min\{n > 1 :\ t_n \le t_1\}$. Since the random variables $V(H)$ have continuous distributions, the a.s. finiteness of θ_H implies the inequalities $t_{\theta_H} < t_1$ a.s. and $C_{t_1} := \mathbf{E}\, t_{\theta_H}/t_1 < 1$. Let

$$E_{\theta_H} = \mathbf{E}\Bigg(\sum_1^{\theta_H} t_i\Bigg).$$

We prove that the uniform integrability (12) holds if the latter expectation is finite for $H = 1$.

Lemma 4. *Condition* (12) *with* $\varepsilon = \lambda_2^{-1}$ *is necessary and sufficient for* $E_{\theta_1} < \infty$.

Proof. Since $\theta_1 \le \varkappa_1$, the condition (12) implies finiteness of E_{θ_1}. We prove now the converse statement.

First, extend naturally the definition of $\varkappa_H$ onto all positive H. Then, due to the self-similarity of the fluid limits, for any $H > 0$ and any integer n, the random variables $\varkappa_{\lambda_2^n H}$ and $\lambda_2^n \varkappa_H$ are identically distributed.

Second, the embedded Markov chains $\{t_{n+1} = t_n V_n(t_n) K\}$ satisfy the following monotonicity property: if the first chain $t_n^{H_1}$ starts from the state $t_1^{H_1} = H_1$ and the second $t_n^{H_2}$ from the state $H_2 < H_1$, then $t_n^{H_1} \ge t_n^{H_2}$ a.s. for all n.

Third, (11) implies (see also Remark 2 for the case of conditions (**SA1**)) that, for all $H > 0$ and $t \in [1,\lambda_2]$,

$$f_{V(H)}(t) \ge C^0 := \alpha^2 \exp\{-\alpha^2 \lambda_2^2\} > 0. \tag{13}$$

Further, if $K \le \lambda_2^{-1}$, then $V(H)K < 1$ a.s. for any H; therefore, $t_{n+1} < t_n$ a.s. for all n. Hence, $\theta_1 = 2$ and $E_{\theta_1} \le 2$. Additionally, $\mathbf{P}(t_{n+1} \le \lambda_2^{-1/2} t_n \mid t_n) \ge (\sqrt{\lambda_2} - 1)C^0 := p$ a.s. Thus, for any positive H, the random variable μ_H is stochastically dominated by the sum of two independent geometrically distributed random variables with the success probability p; in particular, the random variables $\{\mu_H, H \in [1, \lambda_2]\}$ are uniformly integrable. Thus, the random variables

$$\varkappa_H \le H\mu_H$$

are also uniformly integrable in $H \in [1, \lambda_2]$.

Let now $K > \lambda_2^{-1}$. Then, from (13), there exist $\delta_1, \delta_2 > 0$ such that

$$\mathbf{P}(t_{n+1} \ge (1+\delta_1)t_n \mid t_n) \ge \delta_2 \quad \text{a.s.}$$

Let N be such that $(1+\delta_1)^N \ge \lambda_2^2$. Then

$$\begin{aligned} E_{\theta_1} &\ge \mathbf{E}\left(\sum_1^{\theta_1} t_i \mathbf{I}(t_{k+1} \ge (1+\delta_1)t_k,\ k = 1, \ldots, N-1)\right) \\ &\ge \mathbf{E}\left(\sum_N^{\theta_1} t_i \mathbf{I}(t_{k+1} \ge (1+\delta_1)t_k,\ k = 1, \ldots, N-1)\right). \end{aligned}$$

By the monotonicity, the Markov chain admits a minorizing chain which starts at moment N from the state λ_2^2. Thus, the right-hand side of the latter inequality cannot be smaller than ${\delta_2}^N\, \mathbf{E}\, \varkappa_{\lambda_2^2}(\lambda_2^{-2})$. Since the distributions of the random variables $\varkappa_{\lambda_2 H}(\varepsilon)$ and $\lambda_2 \varkappa_H(\varepsilon)$ are the same for all $\varepsilon > 0$, the mean $\mathbf{E}\, \varkappa_{\lambda_2}(\lambda_2^{-2})$ is finite.

Again, from the monotonicity, for any $H \in [1, \lambda_2]$,

$$\varkappa_H(\lambda_2^{-1}) \le \varkappa_{\lambda_2}(\lambda_2^{-2}) \quad \text{a.s.}$$

The latter inequality implies the uniform integrability of the family $\{\varkappa_H(\lambda_2^{-1}),\ H \in [1, \lambda_2]\}$. △

Lemma 5. (a) *If* $\inf_H \mathbf{E}\, V(H) \ge 1/K$, *then* $E_{\theta_1} = \infty$.

(b) *If* $\sup_H \mathbf{E}\, V(H) < 1/K$, *then* $E_{\theta_1} < \infty$.

Proof. Denote by $\mathcal{F}_i$ the σ-algebra generated by the random variables $\{t_j,\ j \le i\}$. Recall that we consider the initial condition $H = 1 = t_1$.

First, we prove (a). For this, it suffices to consider the case $\theta_1 < \infty$ a.s. Assume to the contrary that $E_{\theta_1} < \infty$. Then

$$\begin{aligned} E_{\theta_1} &= 1 + \sum_{i \ge 2} \mathbf{E}\{t_i \mathbf{I}(\theta_1 \ge i)\} = 1 + \sum_{i \ge 2} \mathbf{E}\,\mathbf{E}\{t_i \mathbf{I}(\theta_1 \ge i) \mid \mathcal{F}_{i-1}\} \\ &\ge 1 + \sum_{i \ge 2} \mathbf{E}\{t_{i-1} \mathbf{I}(\theta_1 \ge i)\} = 1 + \sum_{i \ge 2} \mathbf{E}\{t_{i-1} \mathbf{I}(\theta \ge i-1)\} - C_1 > E_{\theta_1}, \end{aligned}$$

which is a contradiction.

We now prove (b). Let $c := K \sup_H \mathbf{E}\, V(H)$. For any $n \ge 2$,

$$\begin{aligned} E_{\min(\theta_1, n)} &= 1 + \sum_2^n \mathbf{E}\{t_i \mathbf{I}(\theta_1 \ge i)\} \le 1 + c \sum_2^n \mathbf{E}\{t_{i-1} \mathbf{I}(\theta_1 \ge i)\} \\ &\le 1 + c \sum_2^n \mathbf{E}\{t_{i-1} \mathbf{I}(\theta_1 \ge i)\} \le 1 + c \sum_2^{n+1} \mathbf{E}\{t_{i-1} \mathbf{I}(\theta_1 \ge i-1)\} = 1 + cE_{\min(\theta_1, n)}. \end{aligned}$$

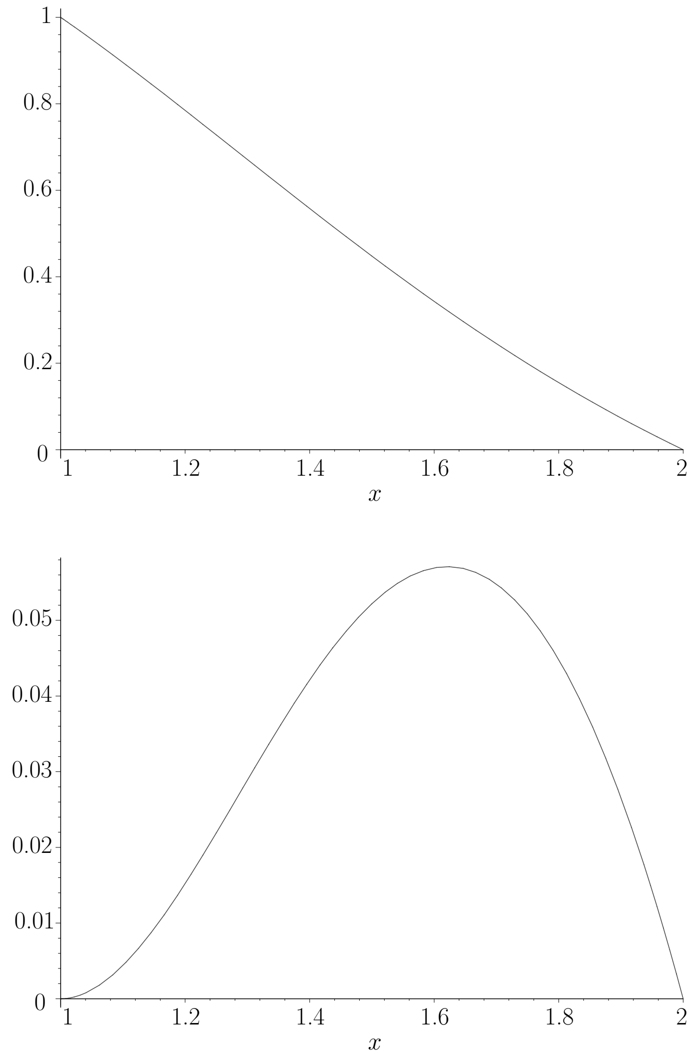


Fig. 3. The graph of equation (14) in the coordinates (λ_2, λ_1) (top) and $(\lambda_2, 2-\lambda_2-\lambda_1)$ (bottom). The ergodicity region corresponds (with a certain accuracy) to the sets below the top graph and above the bottom graph.

Hence,

$$E_{\min(\theta_1,n)} \le \frac{1}{1-c}$$

for any n. Letting n to infinity, we get

$$E_{\theta_1} \le \frac{1}{1-c}. \quad \triangle$$

In the Appendix, we obtain the following estimate.

Theorem 3. *Under assumptions* (**SA2**), *for any* $H \in [1, \lambda_2]$ *we have*

$$\left|\mathbf{E}\, V(H) - \frac{(\lambda_2 - 1)}{\ln \lambda_2}\right| < \frac{2(\lambda_2 - 1)\exp\{-\pi^2 \ln^{-1} \lambda_2\}}{\sqrt{\ln \lambda_2}} \equiv \Delta,$$

where $\Delta = \Delta(\lambda_2) < \dfrac{(\lambda_2 - 1)}{\sqrt{\ln \lambda_2}} \cdot 1.3098 \cdot 10^{-6}$.

We now study conditions for finiteness of the expectations E_{θ_1} in terms of λ_1 and λ_2. From Theorem 3, $\mathbf{E}\,V(H)$ is "almost independent" of H and approximately equal to $(\lambda_2-1)/\ln\lambda_2$.

Figure 3 (top) shows a graph of the solution to equation $K(\lambda_2-1)/\ln\lambda_2 = 1$, which approximately describes the boundary of the regions in Lemma 5. The latter equation has the following form in terms of λ_1 and λ_2:

$$\lambda_1 = \frac{1}{1+\dfrac{(\lambda_2-1)^2}{(2-\lambda_2)\ln\lambda_2}}. \tag{14}$$

In the interval $(1,2)$, this line is close to the straight line $\lambda_1 = 2-\lambda_2$, which corresponds to the necessary stability condition $\lambda_2+\lambda_1 < 2$. The right part of Fig. 3 represents the graph of the difference of the function $\lambda_1 = 2-\lambda_2$ and that given by (14).

We now estimate the difference between the exact boundary of the set from Lemma 5 and the approximate one, which was calculated by the use of formula (14) and is represented in Fig. 3.

The exact boundary lies within the limits determined by the equation $K((\lambda_2-1)/\ln\lambda_2 \pm \Delta(\lambda_2)) = 1$. Denote by $\delta = \delta(\lambda_2)$ the maximum of the absolute value of the difference between the solutions to these equations and equation (14). It is easy to show that $\delta(\lambda_2)$ is "essentially smaller" than $\Delta(\lambda_2)$.

Graphs of the functions $\Delta(\lambda_2)$ and $\delta(\lambda_2)$ are represented in Fig. 4.

We arrive at the following conclusion. From Lemma 5, we have $E_{\theta_1} < \infty$ if the inequality $\sup\limits_H \mathbf{E}\,V(H) < 1/K$ holds, which, in turn, is the case if

$$0 < \lambda_1 < \frac{1}{1+\dfrac{(\lambda_2-1)^2}{(2-\lambda_2)\ln\lambda_2}} - \delta(\lambda_2);$$

also, $E_{\theta_1} = \infty$ if $\inf\limits_H \mathbf{E}\,V(H) \ge 1/K$, which holds if

$$\frac{1}{1+\dfrac{(\lambda_2-1)^2}{(2-\lambda_2)\ln\lambda_2}} + \delta(\lambda_2) < \lambda_1 < 2-\lambda_2,$$

where $1 < \lambda_2 < 2$.

By combining Lemmas 4 and 5 and Theorems 1 and 3, we obtain the following result.

Theorem 4. *Assume that assumptions* (**SA2**) *hold. If* $E_{\theta_1} < \infty$, *then the two-server stochastic system is stable. A sufficient condition for the finiteness of* E_{θ_1} *is* $\sup\limits_H \mathbf{E}\,V(H) < 1/K$. *In turn, the latter inequality holds if*

$$K\left((\lambda_2-1)/\ln\lambda_2 + \Delta(\lambda_2)\right) < 1;$$

the graph of the function $\Delta(\lambda_2)$ *is represented in Fig.* 4, *and* $|\Delta(\lambda_2))| < 1.6\cdot 10^{-6}$.

The latter condition can be rewritten as

$$0 < \lambda_1 < \frac{1}{1+\dfrac{(\lambda_2-1)^2}{(2-\lambda_2)\ln\lambda_2}} - \delta(\lambda_2);$$

the graph of $\delta(\lambda_2)$ *is also represented in Fig.* 4, *and* $|\delta(\lambda_2))| < 3.5\cdot 10^{-8}$.

We continue now our studies of fluid limits in the case $K > \lambda_2^{-1}$. Note that, if $K \ge 1$, then we have $KV(H) > 1$ a.s. for any H; moreover, there exist positive δ_1 and δ_2 such that $\mathbf{P}(KV(H) \ge 1+\delta_1) \ge \delta_2$ for any H. Therefore, $t_n \to \infty$ a.s. if $K \ge 1$.

Assume now that $K \in (\lambda_2^{-1}, 1)$. For any number x, denote by $[x]$ its integer and $\langle x\rangle$ fractional parts. Note that $\langle\langle x\rangle + y\rangle = \langle x+y\rangle$ for any x and y.

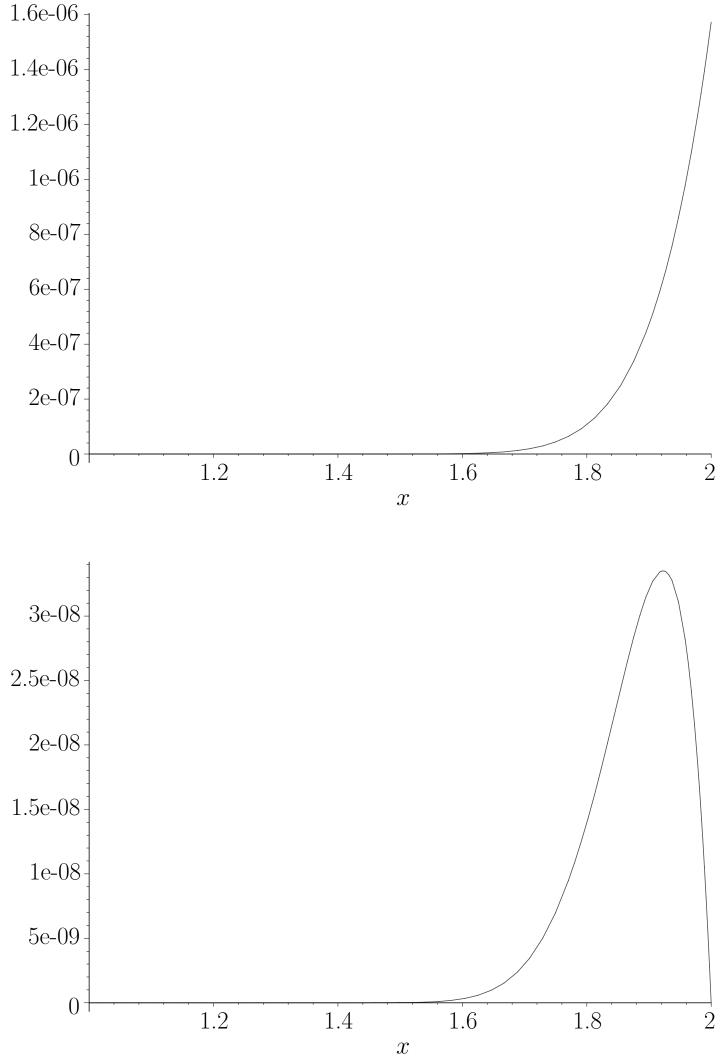


Fig. 4. Graphs of $\Delta(\lambda_2)$ and $\delta(\lambda_2)$.

Lemma 6. *The random variables* $\langle \log_{\lambda_2}(HV(H)) \rangle$ *are identically distributed for all* $H > 0$.

Proof. It suffices to prove the result for $H \in [1, \lambda_2)$. Fix H and let $h = \log_{\lambda_2} H$.
For any $z \in (0,1)$,

$$\begin{aligned}\mathbf{P}\left(\langle \log_{\lambda_2}(HV(H)) \rangle < z\right) &= \mathbf{P}\left(\log_{\lambda_2}(HV(H)) < z\right) + \mathbf{P}\left(1 \le \log_{\lambda_2}(HV(H)) < 1+z\right)\\ &=: P_1(z) + P_2(z).\end{aligned}$$

First, consider the case $0 < z \le h$. Since $V(H) > 1$ a.s., $P_1(z) = 0$. Then,

$$\begin{aligned}P_2(z) &= \mathbf{P}\left(V(H) \ge \frac{\lambda_2}{H}\right) - \mathbf{P}\left(V(H) \ge \frac{\lambda_2^{1+z}}{H}\right)\\ &= \sum_s \left[\mathbf{P}(\lambda_2^{s+1}\alpha \le W_0 < H\lambda_2^{s+1}\alpha) - \mathbf{P}(\lambda_2^{s+1+z}\alpha \le W_0 < H\lambda_2^{s+1}\alpha)\right]\\ &= \sum_s \mathbf{P}(\lambda_2^{s+1}\alpha \le W_0 < \lambda_2^{s+1+z}\alpha) = \sum_s \mathbf{P}(\lambda_2^{s}\alpha \le W_0 < \lambda_2^{s+z}\alpha).\end{aligned}$$

Assume now that $h \le z < 1$. Since $V(H) \le \lambda_2$ a.s., $P_2(z) = \mathbf{P}(HV(H) \ge \lambda_2)$. Therefore,

$$\begin{aligned} P_1(z) + P_2(z) &= 1 - \mathbf{P}\left(\lambda_2^z \le HV(H) < \lambda_2\right) \\ &= 1 - \sum_s \mathbf{P}\left(\lambda_2^{s+z}\alpha \le W_0 < \lambda_2^{s+1}\alpha\right) = \sum_s \mathbf{P}\left(\lambda_2^{s}\alpha \le W_0 < \lambda_2^{s+z}\alpha\right). \end{aligned}$$

Thus, for every $z \in (0,1)$, the value of the sum $P_1(z) + P_2(z)$ is the same for all H. △

Corollary 1. *The random variables* $\langle \log_{\lambda_2}(KHV(H)) \rangle$ *are identically distributed for all* $H > 0$.

Indeed,

$$\langle \log_{\lambda_2}(KHV(H)) \rangle = \langle \langle \log_{\lambda_2}(HV(H)) \rangle + \log_{\lambda_2} K \rangle,$$

where the distribution of the right-hand side does not depend on H.

Since $1 < V(H) < \lambda_2$ a.s. for all H, for $H = 1/K$ we have

$$KHV(H) = V(1/K) \in (1, \lambda_2) \quad \text{a.s.},$$

and therefore $\langle \log_{\lambda_2} V(1/K) \rangle = \log_{\lambda_2} V(1/K)$.

Consider a Markov chain $t_{n+1} = t_n V_n(t_n) K$; let $A_n = [\log_{\lambda_2} t_n]$, $B_n = \langle \log_{\lambda_2} t_n \rangle$, and $H_n = \lambda_2^{B_n}$.

Corollary 2. *For any* $n \ge 2$, *the random variable* B_n *does not depend on the random variables* $t_1, \ldots, t_{n-1}$ *and has the same distribution as the random variable* $\log_{\lambda_2} V(1/K)$.

Note that the distribution of the random variable A_n depends on B_n and $(t_1, \ldots, t_{n-1})$.

We have

$$t_{n+1} = \lambda_2^{A_{n+1}} H_{n+1} = t_n V_n(t_n) K = \lambda_2^{A_n} H_n V_n(H_n) K \quad \text{a.s.}$$

for all n. Hence,

$$B_{n+1} = \log_{\lambda_2} H_{n+1} = \langle \log_{\lambda_2}(H_n V_n(H_n) K) \rangle$$

and

$$A_{n+1} - A_n \overset{\text{a.s.}}{=} [\log_{\lambda_2}(H_n V_n(H_n) K)] =: D_{n+1}. \tag{15}$$

Corollary 3. *Let* $\lambda_2^{-1} < K < 1$. *For any initial value* $t_1 = H$, *the sequence* $\{D_n,\ n \ge 3\}$ *is stationary and ergodic. Additionally,*

$$\mathbf{E}\, D_3 = \mathbf{P}(V_1(1/K) V_2(V_1(1/K)) K \ge \lambda_2) - \mathbf{P}(V_1(1/K) V_2(V_1(1/K)) K < 1), \tag{16}$$

where $\{V_1(H),\ H > 0\}$ *and* $\{V_2(H),\ H > 0\}$ *are two independent copies of the family* $\{V(H),\ H > 0\}$.

Proof. It follows from the construction of the random variables that the distribution of the sequence $\{D_{3+k+n}\}_{n \ge 0}$ is the same for all $k = 0, 1, \ldots$, which this gives the stationarity. Further, since $1 < 1/K < \lambda_2$, there exists a positive ε such that $(1+\varepsilon)/K < \lambda_2$. Additionally, for all $n \ge 2$,

$$\mathbf{P}\left(H_n \ge (1+\varepsilon)/K\right) \ge C\left(1 - \frac{1+\varepsilon}{K\lambda_2}\right) > 0.$$

For $n \ge 2$ and any set $G \subseteq [1, 1+\varepsilon]$,

$$\mathbf{P}\left(H_{n+1} \in G, D_{n+1} = 1 \mid H_n, A_n\right) \ge C\nu(G)$$

a.s. on the event $\{H_n \ge (1+\varepsilon)/K\}$, where $\nu(G)$ is the Lebesgue measure of the set G. Thus, the sequence $\{(H_n, G_n)\}$ is regenerative. Therefore, the sequence $\{G_n\}$ is also regenerative, and its tail σ-algebra is trivial.

Further, the distribution of H_2 coincides with the distribution of the random variable $V_1(1/K)$. It follows from (15) that the random variable D_3 may take only three values: 1, 0, and -1. Therefore,

$$\mathbf{E}\, D_3 = \mathbf{P}(D_3 = 1) - \mathbf{P}(D_3 = -1) = \mathbf{P}(H_2 V_2(H_2) K \geq \lambda_2) - \mathbf{P}(H_2 V_2(H_2) K < 1),$$

and (16) follows. △

Note that $\mathbf{P}(D_3 = -1) = 0$ for $K \geq 1$. We therefore have the following corollary.

Corollary 4. *Let $K > \lambda_2^{-1}$. If $\mathbf{E}\, D_3 < 0$, then $t_n \to 0$ a.s. for any initial value $t_1 = H$. If $\mathbf{E}\, D_3 > 0$, then $t_n \to \infty$ for any $t_1 = H$.*

Proof. Since $1 \leq H_n \leq \lambda_2$ a.s. for all n, we have

$$\lambda_2^{A_n} \leq t_n \leq \lambda_2^{A_n + 1} \quad \text{a.s.}$$

Therefore, $t_n \to \infty$ if and only if $A_n \to \infty$, and $t_n \to 0$ if and only if $A_n \to -\infty$. The proof is completed by applying the SLLN. △

Theorem 5. *Assume that assumptions* (**SA2**) *hold. If $\mathbf{E}\, D_3 > 0$, then the Markov process which describes the two-server queueing system is transient.*

Proof. Note that $\mathbf{P}(D_3 = 1) = 0$ if $K \leq \lambda_2^{-1}$. Therefore, $\mathbf{E}\, D_3 > 0$ implies $K > \lambda_2^{-1}$.

The process $Q(t)$ is strong Markov, and all its states intercommunicate. Let $C > 0$. The number of states with $|Q(t)| \leq C$ is finite for any C. Recall that the process $Q(t)$ is recurrent if (and only if), for any time instant $t \geq 0$, there exists a positive random moment $\gamma(t) \geq t$ a.s. such that $|Q(\gamma(t))| \leq C$ a.s.

We consider the initial state $Q(0) = (C_0, 0, 0)$ for some $C_0 \geq 1$. Denote by T_n the moment of the nth switch of the first server from the second station to the first, and by $X_n = |Q(T_n)|$ the total queue length at time T_n. Here, $X_0 = C_0$. Let $N \geq 1$ be an arbitrary integer. If the process $Q(t)$ is recurrent, then, for some (sufficiently large) C and any time instant $t \geq 0$, one can chose a random number $\nu(t) \geq 1$ such that $T_{N\nu(t)} \geq t$ a.s. and $X_{N\nu(t)} \leq 2C$ a.s.

Below we determine N and show that, under the conditions of the theorem, $X_{Nn} \to \infty$ a.s. as $n \to \infty$, which implies the transience of the process.

Note that, in the fluid model, $t_{n+1} > K t_n > \lambda_2^{-1} t_n$ a.s. for all n. Let $N \geq 3$ be such that

$$\mathbf{E}\left(\sum_{n=3}^{N} D_n\right) > 3 - \log_{\lambda_2}(1 - \lambda_1).$$

Then

$$\inf_{H>0} \mathbf{E}\left(\log_{\lambda_2}(x_N) - \log_{\lambda_2}(x_0) \mid x_0 = (1 - \lambda_1)H\right) =: \gamma > 0.$$

Further, for the underlying stochastic model, one can repeatedly apply Chernoff's inequality and get the following (rough but sufficient) estimates: for certain constants C_1 and C_2,

$$\mathbf{P}(X_1 \geq k X_0 \mid X_0) \leq C_1 \exp\{-C_2 k\} \tag{17}$$

and

$$\mathbf{P}(X_1 \leq X_0/k \mid X_0) \leq C_1 \exp\{-C_2 k\} \tag{18}$$

for all $k \geq 1$ and all $X_0 \geq 1$. We provide here, for example, an outline of the proof of inequality (17).

Let $X_1 = \widetilde{X}_1 + \widehat{X}_1$, where $\widetilde{X}_1 = Q_1(U_1)$ and $\widehat{X}_1$ is the number of customer arrivals at the first station within the time interval (U_1, R_1). Then

$$\mathbf{P}(X_1 > x) \leq \mathbf{P}\left(\widetilde{X}_1 > x/2\right) + \mathbf{P}\left(\widehat{X}_1 > x/2\right).$$

Since the distribution of $\widetilde{X}_1$ can be estimated from above by the stationary distribution of the queue length in the $M/M/1$ queue with input intensity λ_1 and service intensity 1, there exist constants C_3 and C_4 such that

$$\mathbf{P}\left(\widetilde{X}_1 > x/2\right) \le C_3 e^{-C_4 x}$$

for all x. Further,

$$\mathbf{P}\left(\widehat{X}_1 > kX_0\right) \le \mathbf{P}(Z_1 > kX_0) + \mathbf{P}\left(\widehat{X}_1 > kX_0 \mid Z_1 = kX_0\right).$$

Let $S_n = \sum\limits_1^n \tau_i$ where τ_i are i.i.d. exponential random variables with parameter $\lambda_1 < 1$. Then, for sufficiently small $\alpha > 0$,

$$\mathbf{P}\left(\widehat{X}_1 > kX_0 \mid Z_1 = kX_0\right) = \mathbf{P}\left(S_{kX_0} \le ckX_0\right) = \mathbf{P}\left(e^{-\alpha S_{kX_0}} \ge e^{-\alpha kX_0}\right)$$
$$\le e^{\alpha kX_0}\left(\mathbf{E}\, e^{-\alpha\tau_1}\right)^{kX_0} = \left((1+o(1))\,(1+\alpha)\left(1 - \frac{\alpha}{\lambda_1}\right)\right)^{kX_0}.$$

Thus, for an appropriate choice of α, the value of the term in the parentheses is less than 1.

Analogously, we may estimate the probability $\mathbf{P}(Z_1 > kX_0)$ by the sum

$$\mathbf{P}(Q_2(U_1) > c_1kX_0) + \mathbf{P}(Z_1 > kX_0 \mid Q_2(U_1) = c_1kX_0)$$

and show that, for $0 < c_1 < 2 - \lambda_2$, the second term also admits an exponential bound. We may further estimate $\mathbf{P}(Q_2(U_1) > c_1kX_0)$ by using $\mathbf{P}(U_1 > c_2kX_0)$ and then estimate the latter probability by using $\mathbf{P}(T_1 > c_3kX_0)$ with an appropriate choice of constants c_2 and c_3. As a result, we obtain a sum of exponentially decreasing functions and arrive at inequality (17).

Take the test function $L(x) = \max(0, \log_{\lambda_2} x)$. Then, by (17) and (18), the random variables $L(X_{Nn}) - L(X_0)$ and $(L(X_{Nn}) - L(X_0))^2$ are uniformly integrable in X_0.

It follows from Remark 2 that the convergence $U_1^{(n)}/T_1^{(n)} \to H$ is uniform in $H \in (H_1, H_2)$, for any choice of $0 < H_1 < H_2 < \infty$. The uniform integrability of $\{L(X_{Nn}) - L(X_0)\}$ implies also the convergence of the means. Therefore, there exists a number $C \gg 1$ such that

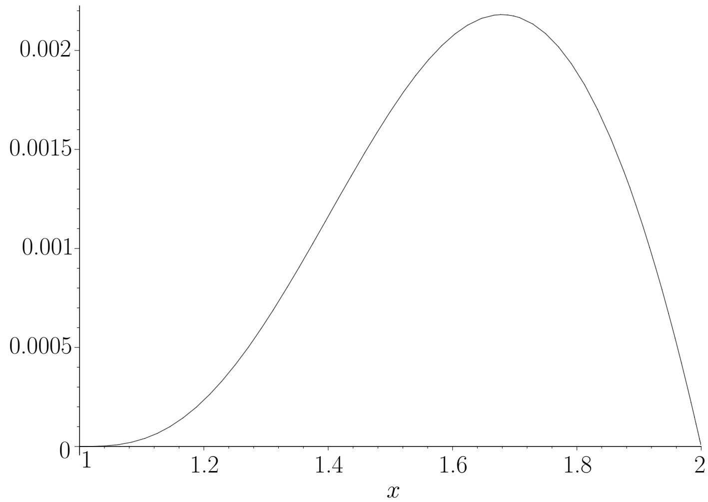
$$\inf_{X_0 \ge C} \mathbf{E}\left(L(X_{Nn}) - L(X_0) \mid X_0\right) \ge \gamma/2.$$

The sequence X_{Nn} forms a time-homogeneous Markov chain, and, for any initial state X_0, there exists an a.s. finite (random) moment n_0 such that $X_{Nn_0} \ge C$ a.s. Thus, the conditions of Theorem 1 from [24] hold, and $L(X_{Nn}) \to \infty$ a.s., which implies the statement of the theorem. △

Remark 4. In the Appendix, we find a Fourier expansion for the density of the random variable $\eta = \log_{\lambda_2} V(H)$. Formula (A.2) provides estimates for coefficients of that expansion, which, in turn, imply that

$$c_0 = 1, \qquad |c_1| + |c_{-1}| < \Delta_0 = 4\pi \exp\left(-\frac{\pi^2}{\ln \lambda_2}\right)\sqrt{\frac{1}{\ln \lambda_2}} < 0.99 \cdot 10^{-5},$$
$$\frac{\sum\limits_{|k| \ge 2} |c_k|}{\Delta_0} < 10^{-4}. \tag{19}$$

Let $\delta_0(\lambda_2) = \sup\limits_H \left|\log_{\lambda_2} V(H) - 1\right|$. From the estimates above, it follows that, uniformly in (H_1, H_2), the joint density of the random vector $\left(\log_{\lambda_2} V_1(H_1), \log_{\lambda_2} V_2(H_2)\right)$ deviates from the uniform density on the unit square by no more than $\delta_0(\lambda_2)$, where $\sup\limits_{\lambda_2} \delta_0(\lambda_2) \le \Delta_0(1{+}10^{-3}) < 10^{-5}$.



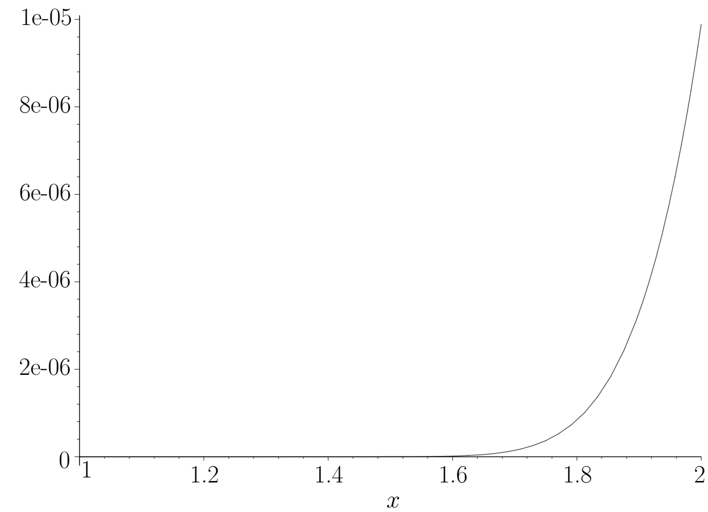


Fig. 5. Top: graph of the difference of the functions (20) and (14) in the coordinates (λ_2, λ_1). Bottom: the graph of the function $\delta_0(\lambda_2)$.

We find now an approximate solution to equation $\mathbf{E}\, D_3 = 0$. Let ζ_1 and ζ_2 be independent random variables uniformly distributed on $[0; 1]$. Then

$$\mathbf{E}\, D_3 \approx \mathbf{P}(\zeta_1 + \zeta_2 + \log_{\lambda_2} K \geq 1) - \mathbf{P}(\zeta_1 + \zeta_2 + \log_{\lambda_2} K < 0).$$

Noting that $s \equiv \log_{\lambda_2}(1/K) \in (0; 1)$, we obtain

$$\mathbf{E}\, D_3 \approx \frac{1}{2}((1-s)^2 - s^2) = \frac{1}{2} - s \approx 0.$$

Finally, we obtain an approximate equation for the boundary: $s = 1/2$, i.e., $K = \lambda_2^{-1/2}$, or, equivalently,

$$\lambda_1 = \frac{\lambda_2^{-1/2}}{\dfrac{\lambda_2 - 1}{2 - \lambda_2} + \lambda_2^{-1/2}}. \tag{20}$$

The graphs of the approximate boundaries of regions of stability and of transience are very close, and we provide in Fig. 5 a graph for the difference of the corresponding functions.

We find a sufficient condition for the inequality $\mathbf{E}\,D_3 > 0$ to hold. From the estimates in Remark 4, $\mathbf{E}\,D_3 > \frac{1}{2} - s - \delta_0(\lambda_2)$. Then $K > \lambda_2^{-1/2+\delta_0(\lambda_2)}$, i.e.,

$$\lambda_1 > \frac{\lambda_2^{-1/2+\delta_0(\lambda_2)}}{\dfrac{\lambda_2 - 1}{2 - \lambda_2} + \lambda_2^{-1/2+\delta_0(\lambda_2)}}.$$

As a result, we obtain the following corollary.

Corollary 5. *Under assumptions* (**SA2**), *a sufficient condition for the instability of the stochastic queueing system is* $K > \lambda_2^{-1/2+\delta_0(\lambda_2)}$, *i.e.*,

$$\lambda_1 > \frac{\lambda_2^{-1/2+\delta_0(\lambda_2)}}{\dfrac{\lambda_2 - 1}{2 - \lambda_2} + \lambda_2^{-1/2+\delta_0(\lambda_2)}},$$

where $|\delta_0(\lambda_2)| < 10^{-5}$ (*see Fig.* 5 (*bottom*) *for the graph of the function* $\delta_0(\lambda_2)$).

Remark 5. The following conjecture seems to be natural: the intermediate case, $E_{\theta_1} = \infty$ and $\mathbf{E}\,D_3 < 0$, corresponds to the null recurrence of the underlying Markov process. However, we are unable to prove this result.

APPENDIX

For brevity, we write $\lambda := \lambda_2$ below. Our proof of Theorem 3 is based on the Fourier expansion of the density of the random variable $\eta = \ln_\lambda V(H)$. These results were obtained with the essential assistance of V.A. Vatutin.

Let $y = \dfrac{\ln(\alpha^2 H)}{\ln \lambda}$. Then, first, $V(H) = \lambda^\eta$ and, second, it is easy to get the following representation of the density of η:

$$f_\eta(x; H) := f_{\eta,y}(x) := \ln \lambda \sum_{k=-\infty}^{\infty} \lambda^k \lambda^{x+y} \exp\{-\lambda^k \lambda^{x+y}\}.$$

This representation is convenient since the function $f_{\eta,y}(x)$ is periodic in the variable $x + y$ with period 1; therefore, an appropriate Fourier expansion may be obtained for the function $f_\eta(x) := f_{\eta,0}(x)$ on the interval $[0, 1]$.

Note that, for any function $g\colon [1, \lambda_2] \to \mathbb{R}$ with finite mean $\mathbf{E}\,g(V(H)) < \infty$, we have the following representation:

$$\mathbf{E}\,g(V(H)) = \mathbf{E}\,g(\lambda^\eta) = \int_0^1 g(\lambda^x) f_{\eta,y}(x)\,dx. \tag{A.1}$$

Lemma 7. *Let*

$$f_\eta(x) = \lambda^x \ln \lambda \sum_{k=-\infty}^{+\infty} \lambda^k e^{-\lambda^{k+x}} := \sum_{k=-\infty}^{\infty} c_k e^{2\pi k i x}.$$

Then

$$c_k = \int_0^1 e^{2\pi k i x} f_\eta(x)\,dx = \Gamma\left(1 + \frac{2\pi k i}{\ln \lambda}\right),$$

where $\Gamma(y)$ *is the standard gamma function for* $y \in \mathbb{C}$.

Proof. By using the identity $e^{2\pi ikx} = e^{2\pi ik(x+n)}$ for $n \in \mathbb{Z}$, we obtain

$$c_k = \ln\lambda \sum_{n=-\infty}^{\infty} \int\limits_0^1 e^{2\pi ikx}\lambda^{x+n}\exp\{-\lambda^{n+x}\}\,dx = \ln\lambda \sum_{n=-\infty}^{\infty} \int\limits_n^{n+1} e^{2\pi ikx}\lambda^{x}\exp\{-\lambda^{x}\}\,dx$$

$$= \ln\lambda \int\limits_{-\infty}^{\infty} e^{2\pi ikx}\lambda^{x}\exp\{-\lambda^{x}\}\,dx = \int\limits_0^{\infty} z^{2\pi ik/\ln\lambda}e^{-z}\,dz = \Gamma\left(1+\frac{2\pi ik}{\ln\lambda}\right). \quad \triangle$$

Proof of Theorem 3. We compute $\mathbf{E}\,V(H)$, using the Fourier expansion of the density from Lemma 7:

$$\mathbf{E}\,V(H) = \int\limits_0^1 \lambda^x \sum_{k=-\infty}^{+\infty} c_k e^{2\pi ki(x+y)}\,dx = \sum_{k=-\infty}^{+\infty} c_k e^{2\pi kiy} \int\limits_0^1 \lambda^x e^{2\pi kix}\,dx$$

$$= (\lambda-1)\sum_{k=-\infty}^{+\infty} \frac{c_k}{\ln\lambda + 2\pi ki} e^{2\pi kiy}.$$

Since $\Gamma(1+x) = x\Gamma(x)$ and $|\Gamma(iy)| = \left(\dfrac{\pi}{y\sinh(\pi y)}\right)^{1/2}$ (see, e.g., [23, p. 43]), for $k \in \mathbb{Z}\setminus 0$ we have

$$b_k := \frac{(\lambda-1)c_k}{\ln\lambda + 2\pi ki}e^{2\pi kiy} = \Gamma\left(\frac{2\pi ik}{\ln\lambda}\right)\frac{1}{\ln\lambda}\frac{2\pi ik(\lambda-1)}{\ln\lambda+2\pi ki}e^{2\pi kiy},$$

$$|b_k| = \left(\frac{\ln\lambda}{2|k|\sinh\dfrac{2\pi^2|k|}{\ln\lambda}}\right)^{1/2}\frac{2|k|(\lambda-1)}{\ln\lambda\sqrt{\ln^2\lambda + (2\pi k)^2}},$$

and, for $k = 0$,

$$b_0 := \frac{(\lambda-1)c_0}{\ln\lambda} = \frac{\lambda-1}{\ln\lambda}.$$

We estimate the coefficients b_k for $k \neq 0$. From the inequality $\sinh(x) \geq e^x/2$, we have, for $x > 0$,

$$|b_k| \leq \left(\frac{\ln\lambda}{|k|}\exp\left(-\frac{2\pi^2|k|}{\ln\lambda}\right)\right)^{1/2}\frac{(\lambda-1)}{\ln\lambda} = \exp\left(-\frac{\pi^2 k}{\ln\lambda}\right)\frac{(\lambda-1)}{\sqrt{k\ln\lambda}},$$

$$q_\lambda := \exp\left(-\frac{\pi^2}{\ln\lambda}\right) \leq \exp\left(-\frac{\pi^2}{\ln 2}\right) := q.$$

By direct calculations, we obtain $q \approx 6.548698489\cdot 10^{-7}$ and

$$2\sum_{k=1}^{\infty} q_\lambda^k = \frac{2q_\lambda}{1-q_\lambda} \leq 2\sum_{k=1}^{\infty} q^k = \frac{2q}{1-q} < 1.3098\cdot 10^{-6}. \quad \triangle$$

We now also obtain estimates for the coefficients c_k. First of all, note that $c_0 = 1$. Analogously to the previous calculations,

$$|c_k| := \left(\frac{\ln\lambda}{2\pi|k|\sinh\dfrac{2\pi^2|k|}{\ln\lambda}}\right)^{1/2}\frac{2\pi k}{\ln\lambda} \leq q_\lambda^{|k|}\sqrt{\frac{|k|}{\ln\lambda}}, \quad k\in\mathbb{Z}\setminus\{0\}, \tag{A.2}$$

and the latter inequality now implies (19) in Remark 4.

We conclude by remarking that, based on the approach of the proof of Theorem 5 and the representation (A.1), one can relatively easily get "good" approximations for a wide class of functions $g(x)$.

The authors express their deep gratitude to Prof. V.A. Vatutin for his interest to and operative assistance with solution of problems discussed in the Appendix, and to Prof. S. Zachary for improving the style of the English translation.

REFERENCES

1. Rybko, A.N. and Stolyar, A.L., Ergodicity of Stochastic Processes Describing the Operation of Open Queueing Networks, *Probl. Peredachi Inf.*, 1992, vol. 28, no. 3, pp. 3–26 [*Probl. Inf. Trans.* (Engl. Transl.), 1992, vol. 28, no. 3, pp. 199–220].
2. Dai, J.G., On Positive Harris Recurrence of Multiclass Queueing Networks: A Unified Approach via Fluid Models, *Ann. Appl. Probab.*, 1995, vol. 5, no. 1, pp. 49–77.
3. Stolyar, A.L., On the Stability of Multiclass Queueing Networks: A Relaxed Sufficient Condition via Limiting Fluid Processes, *Markov Process. Related Fields*, 1995, vol. 1, no. 4, pp. 491–512.
4. Dai, J.G. and Meyn, S.P., Stability and Convergence of Moments for Multiclass Queueing Networks via Fluid Limit Models, *IEEE Trans. Automat. Control*, 1995, vol. 40, no. 11, pp. 1889–1904.
5. Chen, H., Fluid Approximations and Stability of Multiclass Queueing Networks: Work-Conserving Disciplines, *Ann. Appl. Probab.*, 1995, vol. 5, no. 3, pp. 637–665.
6. Down, D.G., On the Stability of Polling Models with Multiple Servers, *CWI Report BS-R 9605*, Amsterdam, 1996.
7. Foss, S.G. and Rybko, A.N., Stability of Multiclass Jakson-Type Networks, *Markov Process. Related Fields*, 1996, vol. 2, no. 3, pp. 461–486.
8. Meyn, S.P., Transience of Multiclass Queueing Networks and Their Fluid Models, *Ann. Appl. Probab.*, 1995, vol. 5, no. 4, pp. 946–957.
9. Puhalskii, A.A. and Rybko, A.N., Nonergodicity of a Queueing Network under Nonstability of Its Fluid Model, *Probl. Peredachi Inf.*, 2000, vol. 36, no. 1, pp. 26–47 [*Probl. Inf. Trans.* (Engl. Transl.), 2000, vol. 36, no. 1, pp. 23–41].
10. Kumar, P.R. and Seidman, T.I., Dynamic Instabilities and Stabilization Methods in Distributed Real-Time Scheduling of Manufacturing Systems, *IEEE Trans. Automat. Control*, 1990, vol. 35, no. 3, pp. 289–298.
11. Chen, H. and Mandelbaum, A., Discrete Flow Networks: Bottleneck Analysis and Fluid Approximations, *Math. Oper. Res.*, 1991, vol. 16, no. 2, pp. 408–446.
12. Bramson, M., Instability of FIFO Queueing Networks, *Ann. Appl. Probab.*, 1994, vol. 4, no. 2, pp. 414–431.
13. Malyshev, V.A., Networks and Dynamical Systems, *Adv. Appl. Probab.*, 1993, vol. 25, no. 1, pp. 140–175.
14. Ignatyik, I. and Malyshev, V., Classification of Random Walks in $\mathbb{Z}_+^4$, *Selecta Math.*, 1993, vol. 12, no. 2, pp. 129–166.
15. Fayolle, G., Malyshev, V.A., and Menshikov, M.V., *Topics in the Constructive Theory of Countable Markov Chains*, Cambridge: Cambridge Univ. Press, 1995.
16. Foss, S. and Kovalevskii, A., A Stability Criterion via Fluid Limits and Its Application to a Polling System, *Queueing Syst.*, 1999, vol. 32, no. 1–3, pp. 131–168.
17. Hunt, P.J., Pathological Behavior in Loss Networks, *J. Appl. Probab.*, 1995, vol. 32, no. 2, pp. 519–533.
18. Harris, T.E., *The Theory of Branching Processes*, Berlin: Springer, 1963. Translated under the title *Teoriya vetvyashchikhsya sluchainykh protsessov*, Moscow: Mir, 1966.

19. Asmussen, S. and Hering, H., *Branching Processes*, Boston: Birkhäuser, 1983.

20. Athreya, K.B. and Ney, P.F., *Branching Processes*, Berlin: Springer, 1972.

21. Roesler, U., Topchii, V.A., and Vatutin, V.A., The Rate of Convergence for Weighted Branching Processes, *Siberian Adv. Math.*, 2002, vol. 12, no. 4, pp. 57–82.

22. Shiryaev, A.N., *Veroyatnost'*, Moscow: Nauka, 1989, 2nd ed. Translated under the title *Probability*, New York: Springer, 1996.

23. Kratzer, A. and Franz, W., *Transzendente Funktionen*, Leipzig: Akademische, 1960. Translated under the title *Transtsendentnye funktsii*, Moscow: Inostr. Lit., 1963.

24. Foss, S.G. and Denisov, D.E., On Transience Conditions for Markov Chains, *Sibirsk. Mat. Zh.*, 2001, vol. 42, no. 2, pp. 425–433 [*Siberian Math. J.* (Engl. Transl.), 2001, vol. 42, no. 2, pp. 364–371].