S. A. Grishechkin, M. Devetsikiotis, I. Lambadaris, Ch. Hobbs,

DOI: https://doi.org/10.4213/tvp3527

Use of the all-Russian mathematical portal Math-Net.Ru implies that you have read and agreed to these terms of use
http://www.mathnet.ru/eng/agreement

Download details:
IP: 54.70.40.11
April 18, 2019, 01:53:31
The last formula is a simple consequence of the relation \( E e^{-s u(t)} = \exp(-sx/(1+st/2)), \) 
\( s > -2/t \) (see, e.g., [4]).

In conclusion note that the one-dimensional distributions of the process \( b(t) \) are of the same type as those of the process \( \eta \); the general finite-dimensional distributions are, of course, quite different (for more detail see, e.g., [3]).

REFERENCES

2. Боровков К. А. О скорости сходимости ветвящихся процессов к диффузионному. — Теория вероятности и ее применения, 1985, т. 30, в. 3, с. 468–477.
10. Колмогоров А. Н. Переход ветвящихся процессов в диффузионные и примыкающие задачи генетики. — Теория вероятности и ее применения, 1959, т. 4, в. 2, с. 233–236.
12. Лебедев Е. А. Предельные теоремы для критических ветвящихся процессов в случайной среде. — Теория вероятности и матем. статист., 1979, т. 20, с. 68–76.
Multistability in queues

Introduction. Although the main mathematical result of this paper is actually about the «large deviation» type of behavior of subcritical Markov multitype branching processes with infinite variance, this topic is hidden inside proofs and we begin to discuss the problem that looks completely different. Branching processes will appear later on.

It is a common situation in queueing theory that a system is roughly speaking either recurrent with a unique steady-state regime or transient. It has been observed by NORTEL (Northern Telecom) that sometimes communication networks behave in a manner not in accordance with either of these types. Besides the main steady-state regime it appears that the system has another operating regime with inferior performance corresponding to much bigger queue sizes on network switches. This second state is not true congestion in the sense that the system does not tend to fall down immediately as it should according to the transient scenario. All characteristics stabilize for a time as we deal with another stationary point of the system.

The investigation of this phenomenon, so-called multistability, is difficult since potentially it can be initiated by a number of reasons. In our paper we construct a simple model that exhibit such a behavior due to one factor: retransmission. Retransmission usually arises if the system is very busy and transmission time of a cell exceeds some specified threshold. Then it is assumed that this cell is lost and retransmission occur. It is the essence of common retransmission protocols (like back-up-N, see [10]).

Current networks transmit at the same time different types of information. It is important to note that not all cells need to be retransmitted. For example, real-time applications, like video, are not retransmissible simply because it is usually too late. On the contrary, the data, like FTP, is a typical example of retransmissible information. So in our model we assume two types of cells: 1st type (nonretransmissible) and 2nd type (retransmissible).

In our simple model the network will be considered as one black box, i.e., we consider a single-server model. The service discipline is random SIRO. Cells inside the network can pass each other and a random discipline corresponds to the assumption that we have a random mixing of cells inside. In fact simulations show that the change of service discipline does not alter the results significantly, but complicates the analysis due to lack of symmetry.

The arrival streams of both types of cells are assumed to be compound Poisson: batches of cells of i-th type arrive with intensity \( \Lambda_i \). Number of cells in one batch has distribution function \( B_i \), generating function \( b_i(s) \) and the variance of \( B_i \) can be infinite. This assumption is used in the literature to model highly bursty arrivals.

Cells are assumed to be serviced at the end of time slots of duration \( g \). We denote the distribution function of \( g \) by \( G \). Typically, time slots are constant, but we will investigate general \( G \) since it does not complicate the analysis. Denote by \( \rho_i = \Lambda_i(Eg) b_i(1) \) the mean number of cells of i-th type arrived during one time slot. We assume that \( \rho = \rho_1 + \rho_2 < 1 \) which guarantees the existence of the main steady-state regime.

If, at the end of a time slot, the total number \( N_1 + N_2 \) of both types of cells is not less than \( K \), then the system is in retransmission mode at this moment. This means that a retransmissible cell served during this time slot does not leave the system, but returns immediately to the input queue. No new retransmissible cells are accepted by the system when it is in retransmission mode since otherwise the system becomes transient.

Although we have found no papers that have considered the problem of retransmission and multistability, the effect of multistability in queueing systems has been previously addressed, see, for example, [6], [7], [10].
2. Results. Eventually the 2-dimensional path of the system (we count the number of cells of both types) hits the border of retransmission \( N_1 + N_2 = K \) (say at point \( B_1 \), Fig. 1). Our first result is that the border can be divided into two areas: stable \((B_3, A_1)\) and unstable \((A_2, B_3)\). In the stable area the process most probably returns to the inside of the triangle \( N_1 + N_2 < K \) after a small delay. Behavior is totally different in the unstable area. Typically the process will go outside the triangle \( N_1 + N_2 < K \) and although it eventually returns, the time taken for this return is very large. This is because there exists another set of «attractors» outside the triangle and the process has a tendency to spend a lot of time in the neighborhood of these attractors. This additional attractor is located at the point \( B_2 \), so all these attractors fill out the axle \( B_3, B_2 \).

This result becomes intuitively clear if we consider the main drift of the system. It is easy to see that this drift is directed always inside the triangle in the stable area (below the horizontal line \( Y = K(1 - \rho_1) \) that passes through the point \( S_3 \)) but towards the line \((B_3, B_2)\) in the unstable area. We omit more detailed arguments and state only a couple of results. Denote by \( T(N_1, N_2) \) the time of return inside the triangle given that the starting point is located at the border \( N_1 + N_2 = K \). It can be proved that

\[
E\tau(N_1, N_2) = (1 - \rho_1)^{-1} \mathbb{E} g \left\{ 1 + N_2 \int_0^1 \frac{z^{N_1}(z^{-1} - 1)}{H(z) - z} \exp \left( N_2 \int_0^s \frac{H(s) - 1}{H(s) - s} ds \right) \frac{dz}{z} \right\},
\]

where \( H(s) = \int_0^\infty \exp(\Lambda_1 u(b_1(s) - 1)) dG(u) \) is the generating function of new arrivals of the 1st type cells during one time slot. An application of the Laplace method demonstrates that this expectation is bounded if \( K \to \infty \) in the stable area and increases exponentially fast in the unstable area.

So it becomes an important question at what area stable or unstable the path of systems hits the border of retransmission because only the unstable area is dangerous to network performance. Our goal is to demonstrate that the tail behavior of \( B_1 \) now becomes of crucial importance. This part of the investigation is motivated by the strong evidence that real traffic in ATM networks is very bursty and can be well described by self-similar models [5], [11]. The model with infinite variance of \( B_1 \) is useful in investigation of such type of traffic (see, for example, [8], [9]).

If the arrival process is ordinary Poisson then it can be shown that the distribution of the first point along the line \( N_1 + N_2 = K \) is a binomial distribution with mean \( EN_i = \frac{K\Lambda_i}{(\Lambda_1 + \Lambda_2)} \) and the peak near the point \( B_4 \) (see Fig. 1). We see that in this
case the most probable first hit point is always located inside the stable area. Moreover,
since $K$ may reasonably be assumed to be large we see that the probability of the process
hitting the border of retransmission in the unstable area is extremely small. Apparently
this conclusion is valid not only when $b_i(s) = s, i = 1, 2$, but for other distributions $B_i$
with rapidly decreasing tails.

The situation is quite different if $B_i$ has infinite variance. Because a mathematical
analysis is very cumbersome we consider only the purely symmetrical case $\Lambda_1 = \Lambda_2 = \Lambda,
B_1 = B_2 = B$. Because of the symmetry, the distribution of the first hit point along the line
$X + Y = K$ should be symmetrical with respect to the middle of this interval and we saw
that in the Poisson case almost the whole mass is concentrated in the middle. Simulations
show an interesting effect: in the case when $B$ has infinite variance, the distribution is
concentrated not in the middle but near the ends of the interval, as in the arcsine law so
the probability of hitting the «bad» area is near $|$. Note that since the first hit point does
not depend on the behavior of the system outside the triangle with border $X + Y = K$,
we can consider for this particular purpose a queueing system ignoring retransmissions.

From the mathematical point of view, an exhaustive analysis of the first hit prob­
lem seems to go beyond current knowledge since we are dealing with a nonhomogeneous
random walk in two-dimensional space. In addition, the infinite variance assumption is
incompatible with any type of diffusion approximation. Instead of investigation of the first
hit point let us take into account not only the first intersection of the path with the border
$X + Y = K$, but all of them; from within and from without the triangle, and consider the
stationary distribution of these points. Despite the distinct natures of these two problems,
we can expect similar shapes of the distributions for the following reason: since $\rho < 1$
the mean drift is towards the inside of the triangle and the system always has a tendency
to return to the origin $(0,0)$. So most probably two successive points of intersection (out
of and into the triangle) are not far from each other. This means that the shapes of the
distributions under consideration should be similar. Even in the case, when the inward
intersection is far from outward one, we can still demonstrate the considered effect using
this approach, since the outward intersections alone should make the shape of the density
$\mathcal{U}$-similar.

To investigate all intersections it is necessary roughly speaking, to consider the state
of the system at all points of the path and find the stationary distribution of these points.
The difficulty is that a batch of cells could arrive at one moment but the path considered
has consecutive increment/decrements in steps of one cell. To incorporate this feature, we
introduce the stretch process $Q(i) = (Q_1(i), Q_2(i)), i = 0, 1, \ldots$, which takes into account
multiple arrivals. For example: suppose there are 3 cells of the first type (at the beginning
of a time slot) and 4 of the second one. Assume that, during this time slot, 3 cells of the
first type arrive in a batch and a second type cell is served. Then the consecutive states of
the stretch process are $(3,4); (4,4); (5,4); (6,4); (6,3)$ (we assume that a service is completed
at the end of a time slot). Our aims are: (a) to find a steady-state distribution of the
stretch process and (b) to investigate its limiting behavior along a line of type $X + Y = K$
as $K \to \infty$.

Denote by $\Pi(\lambda_1, \lambda_2)$ the Laplace transform of the number of cells in the stretch
process in steady state and let $S(dX, dY)$ be the corresponding measure. For investigation
of $\Pi(\lambda_1, \lambda_2)$ we use the branching process technique. Consider a Markov branching process
with continuous time and two types of particles. Each particle (independent of type) lives
an exponentially distributed time with mean 1 and then splits into a random number of
descendants in accordance with generating function

$$H(s_1, s_2) = \int_0^\infty \exp(\Lambda u(b(s_1) + b(s_2) - 2)) dG(u)$$

(this being the number of new cells arriving during one time slot). Denote by $z(t) = (z_1(t),
z_2(t))$ the number of particles at time $t$ in this branching process. Then the Laplace
transform

$$F_i(t, \lambda_1, \lambda_2) = \mathbb{E}\left \{ \exp(-\lambda_1 z_1(t) - \lambda_2 z_2(t)) \mid z_j(0) = \delta_j \right \}, \quad i = 1, 2,$$
satisfies the following equations (see [1]):

$$\frac{\partial}{\partial t} F_i(t, \lambda_1, \lambda_2) = H(F_1, F_2) - F_i, \quad F_i(0, \lambda_1, \lambda_2) = e^{-\lambda_i}, \quad i = 1, 2.$$  (2)

The steady-state distribution $\Pi(\lambda_1, \lambda_2)$ is connected with $F_i(t, \lambda_1, \lambda_2)$ in the following way.

**Theorem 1.** We have

$$\Pi(\lambda_1, \lambda_2) = \text{const} \left[ 1 + G(\lambda_1, \lambda_2) \left\{ \frac{\partial}{\partial \lambda_1} I(\lambda_1, \lambda_2) + \frac{\partial}{\partial \lambda_2} I(\lambda_1, \lambda_2) \right\} \right],$$  (3)

where

$$G(\lambda_1, \lambda_2) = \int_0^\infty \left\{ 1 + \frac{e^{-\lambda_1 e^{-\lambda_2}}}{1 - e^{-\lambda_1 e^{-\lambda_2}}} \right\} dG(u),$$

$$I(\lambda_1, \lambda_2) = \int_0^\infty \left\{ 1 - 2H(F_1(t, \lambda_1, \lambda_2), F_2(t, \lambda_1, \lambda_2)) \right\} dt,$$

$$b(s_1, s_2) = \frac{1}{2} (b(s_1) + b(s_2)), \quad \overline{b}(s_1, s_2) = \frac{1}{2} (\overline{b}(s_1) + \overline{b}(s_2)), \quad \overline{s}(s) = \frac{sb(s) - s}{s - 1}$$

and \text{const} is a normalizing constant.

The proof of this theorem is given below.

We will investigate the limiting behavior of the measure $S$ under the plausible assumption that the distribution $G$ has a rapidly decreasing tail (typically it is constant) and $B$ has infinite variance and a tail of Pareto type $1 - B(u) = u^{-\alpha} L(1/u), \ 1 < \alpha < 2$, where $L(1/u) \rightarrow L \in (0, \infty)$ as $u \rightarrow \infty$ (the use of arbitrary slowly varying function $L$ complicates the proof and is uncommon in queueing theory). Let $S, \{A\} = \gamma \gamma^\alpha S \{\gamma^{-1} A\}$, where the parameter $\gamma \rightarrow 0$. These measures describe the tail of $S$. We want to prove that there exists a measure $\mu$ such that $S, \{A\} \Rightarrow \mu$ in $R^2 \setminus \{(0, 0)\}$. This would mean that the behavior of $S(dx, dy)$ along the line of type $x + y = m, x \neq 0, y \neq 0$ could be approximated by the behavior of $\mu(dx, dy)$ along the line $x + y = k$. This is the type of assertion for which we are looking. Note that, because of the normalization $\gamma^{-1-\alpha}$, measure $\mu$ cannot be defined for any set which contains $(0, 0)$.

Measure $\mu$ can most conveniently be described in coordinates $r = (x + y)/\sqrt{2}, z = (x - y)/(x + y)$. These coordinates can be interpreted as follows: let a point $A = (x_0, y_0)$ be on the line $x + y = m, x \geq 0, y \geq 0$. Then $r$ is the perpendicular distance between this line and the origin, so all points on the line $x + y = m, x \geq 0, y \geq 0$ have the same $r$. The midpoint $M = (m/2, m/2)$ of this line has $z = 0$ and the $x$ coordinate of $A$ is $|A - M|/r$, where the positive sign corresponds to $0 < \varphi < \pi/4$ and the negative sign corresponds to $\pi/4 < \varphi < \pi/2$. In particular edge points $(m, 0)$ and $(0, m)$ have $z = \pm 1$.

**Theorem 2.** $S, \{A\} \Rightarrow \mu$ in $R^2 \setminus \{(0, 0)\}$ as $\gamma \rightarrow 0$. The limiting measure $\mu = \mu^a + \mu^s$ is a sum of absolutely continuous measure $\mu^a$ and singular measure $\mu^s$ which have the following structure. The measure $\mu^a$ is a sum $\mu^a = \mu^a_1 + \mu^a_2$, where measure $\mu^a_1$ is concentrated on the axis $OX$, measure $\mu^a_2$ is concentrated on the axis $OY$, and

$$\mu^a_1 \left\{ [u, \infty) \times [0, \infty) \right\} = \mu^a_2 \left\{ [0, \infty) \times [u, \infty) \right\} = C_1 u^{1-\alpha}, \quad u > 0.$$

Measure $\mu^a$ has a density

$$q(r, z) = C_2 r^{-\alpha} |z|^{(\alpha/\rho) - 1 - \alpha}, \quad r > 0, \ -1 < z < 1,$$

where

$$C_1 = \text{const} \frac{\rho \Lambda(Eg) L}{(1 - \rho)(\alpha - 1)}, \quad C_2 = \text{const} \frac{(1 + \rho) \alpha 2^{(\alpha - 1)/2} \Lambda(Eg) L}{\rho(1 - \rho)}.$$

We are interested in the behavior of $\mu$ along the line $x + y = k$. It is given in the following corollary which immediately follows from Theorem 2.
Corollary 1. A conditional distribution of the measure $\mu$ along the line $x + y = k$ has the following structure. It has two atoms $C_3$ in the edge points $z = \pm 1$ and inside the line the distribution is absolutely continuous with the density $q(z) = C_4 |z|^{-(1 + \alpha) + \alpha/\rho}$, $-1 < z < 1$, where

$$C_3 = \frac{C_1(\alpha - 1)}{\sqrt{2}} C_1(\alpha - 1)/\sqrt{2 + 2C_2/\alpha(1 - \rho)},$$

$$C_4 = \frac{2C_2\rho/\alpha(1 - \rho)}{C_1(\alpha - 1)/\sqrt{2 + 2C_2\rho/\alpha(1 - \rho)}}.$$

Let us note one interesting effect: there are two quite different shapes of the density $|z|^{-(1 + \alpha) + \alpha/\rho}$. If $\alpha\{\rho^{-1} - 1\} > 1$ then this density has maximum at $z = \pm 1$ and equals 0 at $z = 0$, but if $\alpha\{\rho^{-1} - 1\} < 1$ then the density has minimum at $z = \pm 1$ and tends to infinity as $z \to 0$.

We try now to give an intuitive explanation of Corollary 1 since the proof below is rather technical and does not provide enough intuition.

The mechanisms of outward and inward intersections are quite different. Outward intersections are caused by large deviations. They arise when a very large batch of cells arrives. Since any batch contains cells only of one type we have a very large jump along the direction of axis infinity as $z \to 0$. This system of differential equations $u(t) = \rho/2 - u/(x + y)$, $v(t) = \rho/2 - y/(x + y)$. If $\rho$ is small, the path returns promptly to the origin and the inward intersection will not be far from $(0, k)$. But if $\rho$ is close to 1 then the path will spend a considerable time outside triangle. Because of the form of the mean drift, the path is always attracted to the diagonal $x + y = k$ and this forms the peak of the density $q(z)$ near $z = 0$ if $\alpha\{\rho^{-1} - 1\} < 1$. This reasoning can be made more rigorous: let us neglect random fluctuations and assume that the return path is directed strictly along the vector $a$. In other words the path is the solution of the system of differential equations $x(t) = \rho/2 - x/(x + y)$, $y(t) = \rho/2 - y/(x + y)$. This system could be solved by the substitution $u = x + y$, $v = x - y$ which gives $u(t) = u(0) + (\rho - 1)t$, $v(t) = v(0)u(t)/u(0)]^{1/(1 - \rho)}$. In particular, if the path begins from the point $(J, 0)$, it crosses the border $x + y = k$ at the point $v = k^{1/(1 - \rho)}J^{\rho/(1 - \rho)}$ which corresponds to the $x$-coordinate $x = v/u = (k/J)^{\rho/(1 - \rho)}$. Since the distribution of $J$ is known, we get the distribution of $x$ (i.e., the distribution of the inward intersection)

$$P(z < w) = P((k/J)^{\rho/(1 - \rho)} < w | J > k) = w^{\alpha(1 - \rho)/\rho}, 0 < w < 1.$$
Proof of Theorem 1. Let $M$ be the first epoch when $Q(M) = (0,0)$. This is the regenerative cycle of the stretch process and it is easy to see that $EM < \infty$. For technical convenience we assume that if no cells arrive during the first time slot, then $M = 1$. Let $m(i), m(0) = 0$, be those terms in $Q(\cdot)$ which correspond to the epoch $r(i)$ of the end of the $i$-th time slot. Then $Q(m(i))$ is the embedded Markov chain. We have (with the designation $\lambda Q = \lambda_1 Q_1 + \lambda_2 Q_2$):

$$
\Pi(\lambda_1, \lambda_2) = \frac{1}{EM} \mathbb{E} \left[ \sum_{n=0}^{M-1} e^{-\lambda Q(n)} \right] = \frac{1}{EM} \mathbb{E} \sum_{i=0}^{N-1} \sum_{j=m(i)}^{m(i+1)-1} e^{-\lambda Q(j)}
$$

$$
= \frac{1}{EM} \mathbb{E} \left[ \sum_{i=0}^{\infty} 1(i < N) \sum_{j=m(i)}^{m(i+1)-1} e^{-\lambda Q(j)} \right]
$$

$$
= \frac{1}{EM} \mathbb{E} \left[ \sum_{i=0}^{\infty} 1(i < N) \mathbb{E} \left[ \sum_{j=m(i)}^{m(i+1)-1} e^{-\lambda Q(j)} \mid \mathcal{B}_{r(i)} \right] \right],
$$

where $N$ is the duration of the regenerative cycle of the embedded Markov chain, $m(N) = M$ and $\sigma$-algebra $\mathcal{B}_{r(i)}$ describes the evolution of the queueing system up to time $r(i)$. Clearly the random variable $1(i < N)$ is measurable with respect to $\mathcal{B}_{r(i)}$. Introduce the random sequence $\tilde{Q}(\cdot)$ which differs from $Q(\cdot)$ in that it does not take into account services, so $\tilde{Q}(\cdot)$ has only positive increments. Denote by $\tilde{Q}(0) = (0,0), \ldots , \tilde{Q}(\vartheta(t))$ all states of the process $\tilde{Q}(\cdot)$ in the time interval $[0,t]$ and let $G(t, \lambda_1, \lambda_2) = \mathbb{E}[e^{-\lambda Q(0)} + \cdots + e^{-\lambda Q(\vartheta(t))}]$, $Q(t, \lambda_1, \lambda_2) = \mathbb{E} e^{-\lambda \tilde{Q}(\vartheta(t))}$. It is easy to see that

$$
\mathbb{E} \left[ \sum_{j=m(i)}^{m(i+1)-1} e^{-\lambda Q(j)} \mid \mathcal{B}_{r(i)} \right] = e^{-\lambda Q(m(i))} \int_0^\infty G(t, \lambda_1, \lambda_2) \, dG(t)
$$

$$
= e^{-\lambda Q(m(i))} G(\lambda_1, \lambda_2) \quad \text{(say)}.
$$

Hence

$$
\Pi(\lambda_1, \lambda_2) = G(\lambda_1, \lambda_2) \frac{1}{EM} \mathbb{E} \left[ \sum_{i=0}^{\infty} 1(i < N) e^{-\lambda Q(m(i))} \right]. \quad (4)
$$

The function $G(t, \lambda_1, \lambda_2)$ (and hence $G(\lambda_1, \lambda_2)$) can be found as follows. Consider the process $\tilde{Q}(\cdot)$. During a small time $[t,t+\Delta t]$ with probability $2\lambda \Delta t + o(\Delta t)$ a random number $\xi$ of cells with generating function $b(s_1, s_2)$ enters the system and with probability $1 - 2\lambda \Delta t + o(\Delta t)$ no change occurs. Note that

$$
\mathbb{E} \left[ e^{-\lambda \tilde{Q}(\vartheta(t)+1)} + \cdots + e^{-\lambda \tilde{Q}(\vartheta(t)+\xi)} \mid \tilde{Q}(\vartheta(t)) \right] = \frac{1}{2} \left\{ \tilde{b}(e^{-\lambda_1}) + \tilde{b}(e^{-\lambda_2}) \right\} e^{-\lambda \tilde{Q}(\vartheta(t))}
$$

$$
= \tilde{b}(e^{-\lambda_1}, e^{-\lambda_2}) e^{-\lambda \tilde{Q}(\vartheta(t))},
$$

where $\tilde{b}(s) = (sb(s) - s)/(s - 1)$. Therefore

$$
G(t+\Delta t, \lambda_1, \lambda_2) = 2\lambda \Delta t \mathbb{E} \left[ e^{-\lambda \tilde{Q}(0)} + \cdots + e^{-\lambda \tilde{Q}(\vartheta(t))} + \tilde{b}(e^{-\lambda_1}, e^{-\lambda_2}) e^{-\lambda \tilde{Q}(\vartheta(t))} \right]
$$

$$
+ (1 - 2\lambda \Delta t) G(t, \lambda_1, \lambda_2) + o(\Delta t),
$$

which leads to the differential equation

$$
\frac{\partial}{\partial t} G(t, \lambda_1, \lambda_2) = 2\lambda \tilde{b}(e^{-\lambda_1}, e^{-\lambda_2}) Q(t, \lambda_1, \lambda_2), \quad G(0, \lambda_1, \lambda_2) = 1.
$$

Since $Q(t, \lambda_1, \lambda_2) = \exp\{2\lambda t(b(e^{-\lambda_1}, e^{-\lambda_2}) - 1)\}$, we get

$$
G(t, \lambda_1, \lambda_2) = 1 + \frac{\tilde{b}(e^{-\lambda_1}, e^{-\lambda_2})}{1 - \tilde{b}(e^{-\lambda_1}, e^{-\lambda_2})} \left[ 1 - e^{2\lambda t(b(e^{-\lambda_1}, e^{-\lambda_2}) - 1)} \right].
$$


Note that up to a multiplicative constant the mathematical expectation in the right-hand side of (4) is a steady-state distribution of the embedded Markov chain. To find this distribution, we again use the branching process technique. The approach is quite similar to the one described in [4] so we will give only a brief explanation. Introduce a branching process $z(t) = (z_1(t), z_2(t))$ with immigration at zero: if a number of particles in it becomes zero it restarts. We assume that each restart occurs after an exponentially distributed delay with mean 1 and the process begins with a random number of particles with generating function $H(s_1, s_2)$. The basic idea is that this branching process and our queueing system have exactly the same embedded Markov chain. The Laplace transforms of steady-state distributions of the branching process $\Pi z(A_1, A_2)$ and the Markov chain $\Pi m(\lambda_1, \lambda_2)$ are related as follows

$$\Pi m(\lambda_1, \lambda_2) = \epsilon \left\{ \Pi z(\infty, \infty) - \frac{\partial}{\partial \lambda_1} \Pi z(\lambda_1, \lambda_2) - \frac{\partial}{\partial \lambda_2} \Pi z(\lambda_1, \lambda_2) \right\},$$

where $\epsilon$ is a normalizing constant. The proof of this formula is very similar to [4] and is therefore omitted. The Laplace transform $\Pi z(\lambda_1, \lambda_2)$ can be derived by standard regenerative arguments [2]. Let $a$ be a regenerative cycle of the process $z$. We have

$$\Pi z(\infty, \infty) = \frac{1}{\text{E}_a}, \quad 1 - \Pi z(\lambda_1, \lambda_2) = \frac{1}{\text{E}_a} \int_0^\infty \left\{ 1 - H(F_1(t, \lambda_1, \lambda_2), F_2(t, \lambda_1, \lambda_2)) \right\} dt.$$ 

Formula (3) follows from (4), (5). The theorem is proved.

Theorem 2 follows from several lemmata which we now formulate.

**Lemma 1.** Let $\int_0^\infty u^d G(u) du < \infty$ and $1 - B(u) = u^{-\alpha} L(1/u)$, $u \to \infty$, $1 < \alpha < 2$, where $L(1/u) \to L(0, \infty)$. Then

$$1 - b(1 - x) = b (1 - x - x^\alpha L_1(x),$$

$$1 - H(1 - x_1, 1 - x_2) = \frac{\rho x_1}{2} + \frac{\rho x_2}{2} - r(x_1, x_2),$$

$$G(x_1, x_2) = 1 + \rho - L_1 \Lambda(\text{E}_g)(x_1^{\alpha - 1} + x_2^{\alpha - 1}) + o(x_1^{\alpha - 1} + x_2^{\alpha - 1}),$$

where

$$r(x_1, x_2) = x_1^\alpha L_2(x_1) + x_2^\alpha L_2(x_2),$$

$$L_1(x) \to L_1 = L(\alpha - 1)^{-1} \Gamma(2 - \alpha),$$

$$L_2(x) \to L_2 = L\Lambda(\text{E}_g)(\alpha - 1)^{-1} \Gamma(2 - \alpha), \quad x \to 0.$$ 

The proof of this lemma is a standard application of Tauberian theorems [3] and is therefore omitted.

Later we need the second term of the asymptotics of $F_i(t, \lambda_1, \lambda_2)$ as $\lambda_1, \lambda_2 \to 0$. Consider the functions $h(x_1, x_2) = 1 - H(1 - x_1, 1 - x_2), \Psi_i = 1 - F_i$ and rewrite (2) in the following form

$$\frac{\partial}{\partial t} \Psi_i(t, \lambda_1, \lambda_2) = H(\Psi_1, \Psi_2) - \Psi_i, \quad \Psi_i(0, \lambda_1, \lambda_2) = 1 - e^{-\lambda_i}, \quad i = 1, 2.$$ 

Mathematical expectations $A_i(t) = -\frac{\partial}{\partial \lambda_i} F_i(t, 0, 0)$ satisfy the matrix equation

$$\frac{d}{dt} A(t) = A(t) M = MA(t), \quad A(0) = E,$$

where

$$A(t) = \begin{pmatrix} A_1(t) & A_2(t) \\ A_1^2(t) & A_2^2(t) \end{pmatrix}, \quad M = \begin{pmatrix} \frac{\rho_1}{2} - 1 & \frac{\rho_2}{2} \\ \frac{\rho_1}{2} & \frac{\rho_2}{2} - 1 \end{pmatrix}.$$ 

The solution of (9) is

$$A_1(t) = A_2^2(t) = \frac{1}{2} \left( e^{\mu_1 t} + e^{\mu_2 t} \right), \quad A_2(t) = A_1^2(t) = \frac{1}{2} \left( e^{\mu_1 t} - e^{\mu_2 t} \right),$$

where $\mu_1 = -1, \mu_2 = \rho - 1.$
Let $t, \lambda_i' \geq 0$, $i = 1, 2$, be fixed and $\lambda_i = \gamma \lambda_i', \gamma \to 0$. Since

\[ \Psi_i(t, \lambda_1, \lambda_2) = E \{ 1 - \exp(-\lambda_1 z_1(t) - \lambda_2 z_2(t)) \mid z_j(0) = \delta_j \} \]

and $1 - \exp(-\lambda_1 z_1(t) - \lambda_2 z_2(t)) \leq \lambda_1 z_1(t) + \lambda_2 z_2(t)$ we can apply the dominated convergence theorem and get

\[ \gamma^{-1} \Psi_i(t, \lambda_1, \lambda_2) \to A_i^1(t) \lambda'_1 + A_i^2(t) \lambda'_2 = A_i(t, \lambda'_1, \lambda'_2) \quad \text{(say)}, \]

where $\lambda_i = 7 \lambda_i'$, $7 \to 0$. Since

\[ \phi_i(t, \lambda_1, \lambda_2) = \exp \left( -A_i(t) \lambda'_1 - A_i^2(t) \lambda'_2 \right) \]

we can apply the dominated convergence theorem and get

\[ \gamma^{-1} \phi_i(t, \lambda_1, \lambda_2) \to \alpha_i(t) A_i(t) \lambda'_1 + \beta_i(t) A_i^2(t) \lambda'_2. \]

To find the second term of the asymptotic, consider functions $R_i(t, \lambda_1, \lambda_2) = \alpha_i(t) A_i(t) \lambda'_1 + \beta_i(t) A_i^2(t) \lambda'_2 - \phi_i(t, \lambda_1, \lambda_2)$. These functions satisfy the equations

\[ \frac{d}{dt} \left( \begin{array}{c} R_1 \\ R_2 \end{array} \right) = \mathbf{M} \left( \begin{array}{c} R_1 \\ R_2 \end{array} \right) + \left( \begin{array}{c} r_1(\Psi_1, \Psi_2) \\ r_2(\Psi_1, \Psi_2) \end{array} \right) \]

with $R_i(0, \lambda_1, \lambda_2) = -1 + e^{-\lambda_1} + \lambda_i$ and the function $r$ as defined in Lemma 1.

**Lemma 2.** Let $t, \lambda_i' \geq 0$, $i = 1, 2$, be fixed and $\lambda_i = \gamma \lambda_i', \gamma \to 0$. Then

\[ R_i(t, \lambda_1, \lambda_2) \sim \gamma^{-\alpha} L_2 e^{\mu_2 t} \int_0^t e^{-\mu_2 u} \left\{ (A_i^1(u) \lambda'_1 + A_i^2(u) \lambda'_2)^{\alpha} + (A_i^1(u) \lambda'_1 + A_i^2(u) \lambda'_2)^\alpha \right\} du, \quad i = 1, 2. \]

**Proof.** It follows from (12) that

\[ R_i(t, \lambda_1, \lambda_2) = e^{\mu_2 t} \int_0^t e^{-\mu_2 u} \left( \frac{d}{du} \right) \left( \begin{array}{c} R_1 \\ R_2 \end{array} \right) + \left( \begin{array}{c} r_1(\Psi_1, \Psi_2) \\ r_2(\Psi_1, \Psi_2) \end{array} \right) du. \]

To obtain these formulae it is necessary to treat (12) as linear differential equations with $\Psi_1, \Psi_2$ considered as known functions. It follows from (13) that

\[ \gamma^{-\alpha} R_i(t, \lambda_1, \lambda_2) = e^{\mu_2 t} \int_0^t e^{-\mu_2 u} \gamma^{-\alpha} r(\Psi_1(u, \lambda_1, \lambda_2), \Psi_2(u, \lambda_1, \lambda_2)) du + o(1). \]

By (11) and the definition of the function $r$ we have

\[ \gamma^{-\alpha} r(\Psi_1(u, \lambda_1, \lambda_2), \Psi_2(u, \lambda_1, \lambda_2)) \to L_2 \left[ A_i^1(t, \lambda'_1, \lambda'_2) + A_i^2(t, \lambda'_1, \lambda'_2) \right]. \]

To complete the proof we need to apply dominated convergence arguments which follow from the inequality (11).

**Lemma 3.** Let $t, \lambda_i' \geq 0$, $i = 1, 2$, be fixed and $\lambda_i = \gamma \lambda_i', \gamma \to 0$. Then

\[ 1 - \Pi(\lambda_1, \lambda_2) = \gamma^{\alpha-1} \text{const} \]

where

\[ c_1 = \frac{\rho \Gamma(2 - \alpha) \Lambda(E\gamma) L}{1 - \rho \alpha \Gamma(2 - \alpha) \Lambda(E\gamma) L}, \quad c_2 = \frac{(1 + \rho) \alpha \Gamma(2 - \alpha) \Lambda(E\gamma) L}{1 - \rho (\alpha - 1)}. \]

**Proof.** It follows from (7) and Lemma 2 that

\[ U(t, \lambda_1, \lambda_2) = -1 + H \left( F_1(t, \lambda), F_2(t, \lambda) \right) + \frac{1}{2} \rho \left\{ A_1(t, \lambda_1, \lambda_2) + A_2(t, \lambda_1, \lambda_2) \right\} \]

\[ \sim \gamma^{\alpha} \left[ L_2 \left\{ A_i^1(t, \lambda'_1, \lambda'_2) + A_i^2(t, \lambda'_1, \lambda'_2) \right\} \right. \]

\[ + \rho L_2 e^{\mu_2 t} \int_0^t e^{-\mu_2 u} \left\{ A_i^1(u, \lambda'_1, \lambda'_2) + A_i^2(u, \lambda'_1, \lambda'_2) \right\} du \right]. \]
It is easy to see using (11), (13) that we can integrate this relationship and get

\[
\int_0^\infty U(t, \lambda_1, \lambda_2) \, dt = -I(\lambda_1, \lambda_2) + \frac{1}{2(1-\rho)} \rho(\lambda_1 + \lambda_2)
\]

\[
\sim \gamma^\alpha \left[ L_2 \int_0^\infty \left\{ A_1^\alpha(t, \lambda'_1, \lambda'_2) + A_2^\alpha(t, \lambda'_1, \lambda'_2) \right\} \, dt 
+ \rho L_2 \int_0^\infty e^{\mu_2 t} \int_0^t e^{-\mu_2 u} \left\{ A_1^\alpha(u, \lambda'_1, \lambda'_2) + A_2^\alpha(u, \lambda'_1, \lambda'_2) \right\} \, du \, dt \right]
\]

\[
= \gamma^\alpha L_3 \int_0^\infty \left\{ A_1^\alpha(u, \lambda'_1, \lambda'_2) + A_2^\alpha(u, \lambda'_1, \lambda'_2) \right\} \, du,
\]

(15)

where \( L_3 = L_2/(1-\rho) \).

Now we would like to differentiate both sides of (15) with respect to \( \lambda_1 \) in order to get an asymptotic of \( \gamma^\alpha I(\lambda_1, \lambda_2) \) that is needed for (3). To justify this procedure we use the following general property of the Laplace transform: if \( \varphi_n(\lambda_1, \lambda_2) \to \varphi(\lambda_1, \lambda_2) \), \( n \to \infty \), where \( \varphi_n(\lambda_1, \lambda_2) \), \( \varphi(\lambda_1, \lambda_2) \) are Laplace transforms of arbitrary random variables, then \( \frac{\partial}{\partial \lambda_1} \varphi_n(\lambda_1, \lambda_2) \to \frac{\partial}{\partial \lambda_1} \varphi(\lambda_1, \lambda_2) \), \( n \to \infty \). To apply this property rewrite (15) in the form

\[
\exp \left\{ -\gamma^\alpha I(\gamma \lambda'_1, \gamma \lambda'_2) \right\} \exp \left\{ \gamma^\alpha \frac{1}{2(1-\rho)} \rho(\gamma \lambda'_1 + \gamma \lambda'_2) \right\}
\]

\[
= \exp \left\{ L_3 \int_0^\infty \left\{ A_1^\alpha(u, \lambda'_1, \lambda'_2) + A_2^\alpha(u, \lambda'_1, \lambda'_2) \right\} \, du \right\}.
\]

(16)

The left-hand side of (16) is the Laplace transform of a random variable since \( \exp\{-\gamma^\alpha I(\gamma \lambda'_1, \gamma \lambda'_2)\} \) is the Laplace transform of the stationary distribution of sub-critical branching process with immigration and two types of particles [1]. Hence we can differentiate both sides of (16) with respect to \( \lambda'_1 \) which leads after simple manipulations to the desired conclusion

\[
- \frac{\partial}{\partial \lambda_1} I(\lambda_1, \lambda_2) + \frac{1}{2(1-\rho)} \rho \sim \gamma^{(1)} L_3 \frac{\partial}{\partial \lambda_1} \int_0^\infty \left\{ A_1^\alpha(u, \lambda'_1, \lambda'_2) + A_2^\alpha(u, \lambda'_1, \lambda'_2) \right\} \, du.
\]

(17)

Now return to formula (3) and denote

\[
\frac{\partial}{\partial \lambda_1} I(\lambda_1, \lambda_2) + \frac{\partial}{\partial \lambda_2} I(\lambda_1, \lambda_2) = \nabla(\lambda_1, \lambda_2).
\]

Since \( \Pi(0,0) = 1 \) we can rewrite (3) as

\[
1 - \Pi(\lambda_1, \lambda_2) = \operatorname{const} \left[ G(0,0) \nabla(0,0) - G(\lambda_1, \lambda_2) \nabla(\lambda_1, \lambda_2) \right]
\]

\[
= \operatorname{const} \left[ G(0,0) \{ \nabla(0,0) - \nabla(\lambda_1, \lambda_2) \} + \nabla(0,0) \{ G(0,0) - G(\lambda_1, \lambda_2) \} \right]
\]

\[
- \left\{ \nabla(0,0) - \nabla(\lambda_1, \lambda_2) \right\} \left\{ G(0,0) - G(\lambda_1, \lambda_2) \right\}.
\]

The asymptotes of \( \nabla(0,0) - \nabla(\lambda_1, \lambda_2) \) and \( G(0,0) - G(\lambda_1, \lambda_2) \) are given by (17) and (8), respectively, (note that \( \nabla(0,0) = \rho/(1-\rho) \)) and it is easy to see that the term \( \{ \nabla(0,0) - \nabla(\lambda_1, \lambda_2) \} \{ G(0,0) - G(\lambda_1, \lambda_2) \} \) has a smaller order \( \gamma^{2\alpha-2} \). The assertion of the lemma follows from these asymptotes.

To transform the deduced asymptotic of \( \Pi(\lambda_1, \lambda_2) \) into probabilistic assertions we need the following general construction.

Denote by \( \mathbb{R}^2_+ \) the quadrant \( x \geq 0, y \geq 0 \). Let \( \mu \) be a two-dimensional measure on \( \mathbb{R}^2_+ \setminus \{(0,0)\} \). We assume that \( \mu\{[x,\infty) \times [y,\infty)\} \) is finite for all \( x, y \geq 0 \) except possibly the point \( x=0, y=0 \). Define the following transform

\[
T_\mu(\lambda_1, \lambda_2) = \int_0^\infty \int_0^\infty e^{-\lambda_1 x - \lambda_2 y} \mu\{[x,\infty) \times [y,\infty)\} \, dx \, dy.
\]

In other words \( T_\mu(\lambda_1, \lambda_2) \) is the usual Laplace transform of the tail of \( \mu \). We have the following properties of \( T_\mu(\lambda_1, \lambda_2) \):
Lemma 4. If a measure \( \mu \) has the Laplace transform \( L(\lambda_1, \lambda_2) \), \( \lambda_1, \lambda_2 \geq 0 \), then

\[
T(\lambda_1, \lambda_2) = \frac{1}{\lambda_1 \lambda_2} \left[ L(0,0) + L(\lambda_1, \lambda_2) - L(0, \lambda_2) - L(\lambda_1, 0) \right].
\]

Proof. By Fubini’s theorem we have

\[
T(\lambda_1, \lambda_2) = \int_0^\infty \int_0^\infty e^{-\lambda_1 x - \lambda_2 y} \int_x^\infty \int_y^\infty \mu(du, dv) \, dx \, dy
\]

(18)

If \( L(\lambda_1, \lambda_2), \lambda_1, \lambda_2 \geq 0 \), exists then the right-hand side of (18) could be rewritten as

\[
\left( \lambda_1 \lambda_2 \right)^{-1} \left[ L(0,0) + L(\lambda_1, \lambda_2) - L(0, \lambda_2) - L(\lambda_1, 0) \right].
\]

The lemma is proved.

Note that \( T(\lambda_1, \lambda_2) \) can exist even if \( L(\lambda_1, \lambda_2) \) does not. A typical example is a measure \( \mu \) with density \( x^{-3/2} y^{-3/2} \).

Lemma 5. Let measures \( \mu_n, \mu \) on \( \mathbb{R}^2 \setminus \{(0,0)\} \) be such that \( T(\lambda_1, \lambda_2) \to T(\lambda_1, \lambda_2), \lambda_1, \lambda_2 \geq 0 \), and let the function \( \mu(\lambda_1, \lambda_2) \) be continuous in \( \lambda_1, \lambda_2 \geq 0 \), \( \lambda_1, \lambda_2 \geq 0 \), and the function \( \mu(x, y) = \mu([x, \infty) \times [y, \infty)) \) be continuous in \( x > 0, y > 0 \). Then \( \mu_n(x, y) \to \mu(x, y), x, y \to 0, n \to \infty, \) where \( \mu_n(x, y) = \mu_n([x, \infty) \times [y, \infty)) \).

If, additionally, \( \mu_n(x, 0) \to \mu(x, 0), \mu_n(0, y) \to \mu(0, y), x, y > 0, n \to \infty, \) then \( \mu_n \to \mu \) in \( \mathbb{R}^2 \setminus \{(0,0)\} \).

Proof. Denote by \( M_n, M \) measures with densities \( \mu_n(x, y), \mu(x, y) \). The usual convergence theorem for the Laplace transforms gives \( M_n \to M \). Now we need to use the monotonicity of \( \mu_n(x, y) \). We have for all sufficiently small \( \varepsilon > 0 \) that

\[
\mu_n(x, y) \leq \varepsilon^{-2} M_n \left\{ (x, x + \varepsilon) \times (y, y + \varepsilon) \right\} \to \varepsilon^{-2} M \left\{ (x, x + \varepsilon) \times (y, y + \varepsilon) \right\}
\]

and hence \( \limsup \mu_n(x, y) \leq \mu(x + \varepsilon, y + \varepsilon) \) for any \( \varepsilon > 0 \). Similarly \( \liminf \mu_n(x, y) \geq \mu(x - \varepsilon, y - \varepsilon) \). The continuity of \( \mu(x, y) \) implies that \( \mu_n(x, y) \to \mu(x, y), x, y > 0 \). This reasoning is not applicable if either \( x \) or \( y \) is equal to zero and the convergence for this case needs to be checked separately. If it is valid, the convergence \( \mu_n \to \mu \) follows by the definition.

Proof of Theorem 2. To prove the convergence \( S_n \to \mu = \mu^\infty + \mu^s, \gamma \to 0 \), it is enough to demonstrate by Lemma 5 that (note that \( T(\lambda_1, \lambda_2) \equiv 0 \))

\[
T_{S_n}(\lambda_1, \lambda_2) \to T(\lambda_1, \lambda_2), \quad \lambda_1, \lambda_2 \geq 0,
\]

(19)

\[
S_n \left\{ [x, \infty) \times [0, \infty) \right\} \to \mu^\infty \left\{ [x, \infty) \times [0, \infty) \right\} + \mu^s \left\{ [x, \infty) \times [0, \infty) \right\},
\]

(20)

\[
S_n \left\{ [0, \infty) \times [y, \infty) \right\} \to \mu^\infty \left\{ [0, \infty) \times [y, \infty) \right\} + \mu^s \left\{ [0, \infty) \times [y, \infty) \right\}, \quad x, y > 0.
\]

(21)

By Lemmas 3, 4 we get

\[
T_{S_n}(\lambda_1, \lambda_2) \to T(\lambda_1, \lambda_2) + T(\lambda_2, \lambda_1),
\]

(22)

where

\[
T(\lambda_1, \lambda_2) = \text{const} \frac{c_2}{\lambda_1 \lambda_2} \int_0^\infty e^{\mu_2 u} \left[ \frac{1}{\lambda_1 \lambda_2} \left( 1 - e^{-\lambda_1 x} \right) (1 - e^{-\lambda_2 y}) \mu^\infty (dx, dy) \right.
\]

\[
- \left. \left\{ \lambda_1 A_1^2(u) + \lambda_2 A_2^2(u) \right\} \right] du.
\]

(23)

Hence, to justify (19) it is necessary to prove that \( T(\lambda_1, \lambda_2) = T(\lambda_1, \lambda_2) + T(\lambda_2, \lambda_1) \). It follows from (18) and definitions of \( c_i, C_i, i = 1, 2, \) that

\[
T(\lambda_1, \lambda_2) = \int_0^\infty \int_0^\infty \frac{1}{\lambda_1 \lambda_2} \left( 1 - e^{-\lambda_1 x} \right) (1 - e^{-\lambda_2 y}) \mu^\infty (dx, dy)
\]

\[
= \frac{\text{const} \left( \alpha - 1 \right)^2 (\alpha - 1)^2 c_2}{\lambda_1 \lambda_2 \Gamma(2 - \alpha) \rho}
\]

\[
\times \int_0^1 \left| z \right|^{(\alpha / \rho) - 1 - \alpha} \int_0^\infty \left( 1 - e^{-\lambda_1 \rho(z+1)/\sqrt{2}} \right) (1 - e^{-\lambda_2 \rho(z-1)/\sqrt{2}}) \, dz.
\]
Integration by parts of the inner integral and substitution \( z = e^{-\rho u} \), \( z > 0 \); \( z = -e^{-\rho u} \), \( z < 0 \), convert this integral into \( T(\lambda_1, \lambda_2) + T(\lambda_2, \lambda_1) \) that proves (19).

To prove (20) let us note that the left-hand side of this relation could be interpreted as the tail of one-dimensional measure with the Laplace transform \( \gamma^{1-\alpha} \Pi(\gamma \lambda_1, 0) \). Standard application of the Tauberian theorem [3, Chapter 13, (5.22)] gives that

\[
S_{\gamma}\left\{(x, \infty) \times [0, \infty)\right\} = \frac{x^{1-\alpha} \text{const}}{\Gamma(2-\alpha)} \left\{ c_1 + c_2 \int_0^\infty e^{\mu_2 u} \left[ \{A_1^1(u)\}^{\alpha-1} + \{A_2^1(u)\}^{\alpha-1} \right] du \right\}. \tag{24}
\]

Straightforward calculations similar to the above one show that

\[
\mu^x\left\{(x, \infty) \times [0, \infty)\right\} = \frac{x^{1-\alpha} \text{const} c_2}{\Gamma(2-\alpha)} \int_0^\infty e^{\mu_2 u} \left[ \{A_1^1(u)\}^{\alpha-1} + \{A_2^1(u)\}^{\alpha-1} \right] du. \tag{25}
\]

Using the definition of the measure \( \mu^x \), (24) and (25), we see that (20) is valid. Formula (21) is proved in the same way. Theorem 2 is proved.

REFERENCES

3. Феллер В. Введение в теорию вероятностей и ее приложения. Т. 2. М.: Мир, 1984, 752 с.

Поступила в редакцию 18.IV.2000