1

Stability Analysis of Queueing Systems based on Synchronization of the Input and Majorizing Output Flows

Larisa AFANASEVA
Department of Probability, Faculty of Mathematics and Mechanics, Lomonosov Moscow State University, Russia

This chapter is focused on the stability conditions for a multiserver queueing system with heterogeneous servers and a regenerative input flow X. The main idea is constructing an auxiliary service process Y, which is also a regenerative flow and determination of the common points of regeneration for both processes X and Y. Then the traffic rate is defined in terms of the mean of the increments of these processes on a common regeneration period. It allows us to use well-known results from the renewal theory to find the instability and stability conditions. The possibilities of the proposed approach are demonstrated by examples. We also present the applications to transport system capacity analysis.

1.1. Introduction

This chapter deals with the study of stability conditions for a heterogeneous multiserver queueing system with regenerative input flow (Afanaseva 2019).

We consider queueing systems with regenerative input flow for three reasons. First, a process describing the performance of the system under some natural conditions

Queueing Theory 2,
coordinated by Vladimir ANISIMOV, Nikolaos LIMNIOS.

turns out to be a classical regenerative process (Asmussen 2003; Thorisson 2000) and the renewal theory gives very effective tools for asymptotic analysis of the system. Second, the class of regenerative flows is rather wide and includes fundamental flows from queueing theory. Finally, regenerative flow has some useful properties that allow us to investigate various applied models (Afanasyeva and Bashtova 2014).

The main objective of our study is to determine conditions under which the process describing the performance of the system is stochastically bounded and therefore, under some additional assumptions, a stable process.

Let us note that stability results for the classical homogeneous multiserver queue are very well known. We refer to the works of (Kiefer and Wolfowitz 1955) for the $GI|GI|m$ system and the general ergodicity results of (Loynes 1962). As it was shown in Sadowsky (1995), the proposed approach could not be applied for heterogeneous systems. Instead, in the work of (Sadowsky 1995), the stability is examined from the point of view of Harris recurrent Markov chain theory. Based on the works of (Malyshev and Menshikov 1982; Meyn and Tweedie 2009; Georgiadis and Szpankowski 1992; Szpankowski 1994), the author obtained two results concerning the irreducibility and positive Harris recurrence of the corresponding process.

The heterogeneous multiserver queueing system with regenerative flow was investigated in the papers of (Afanasyeva and Tkachenko 2014, 2016, 2018). Under some assumptions, the necessary and sufficient stability condition was obtained there.

We consider m-server queueing system S with regenerative input flow X with intensity λ_X. The main assumption is that the process X does not depend on the stochastic sequences and processes determining the work of the servers, such as service time, the moments of breakdowns and recovery of the servers, and the random numbers of the servers required for the service of the customers.

We construct an auxiliary system S_0 with input flow X_0. The latter does not depend on X and is determined by the stochastic sequence describing the work of the servers. The flow X_0 is constructed in such a way that there are always customers for service. We introduce an auxiliary flow Y, which is the number of customers served in the system S_0 up to time t. Then we establish the conditions of Y being the regenerative flow with the intensity λ_Y.

The basic idea is constructing the common points of regeneration for both processes X and Y. Then we define the traffic rate of the system in terms of the first moments of increments of these processes on the common regeneration period. Under some conditions, we estimate the increments of the service process $\tilde{Y}(t)$ in the original system S (that is the number of customers served in the system up to time t)

with the help of increments of an auxiliary process $Y(t)$. It allows us to prove instability results and to find conditions of stochastic boundedness of the number of customers $Q(t)$ at the system at time t as $t \to \infty$.

One more traditional approach to the analysis of stability of the queueing systems is based on the regenerative structure of the processes describing the states of the system, such as the number of customers in the system or workload process. We make assumptions providing the presence of the points of the regeneration for the processes. Then, it is required to establish the finiteness of the mean of the times between the regenerations in accordance with the theorem of (Smith 1955). It is quite a complex problem for non-Markovian processes. Exactly this approach is employed in several works of Morozov and colleagues (Morozov *et al.* 2011; Morozov 2004, 2007; Morozov and Dimitriou 2017; Morozov and Rumyantsev 2016) for establishing the conditions for stability of a wide circle of models. Although our approach is also based on the notions of regeneration and synchronization, it significantly differs from the approach employed in the works mentioned earlier. Our main objective is to establish necessary and sufficient conditions of the stochastic boundedness of the process, describing the system. Thus, this process does not have to be regenerative. Let us point out that a similar situation emerges in the classic models (Borovkov 1976). The analysis of the stochastic boundedness of the process is important for the applications, since precisely the existence of the stochastic boundaries manifests that the system successfully deals with the task. Our conditions may seem too restrictive to be useful in applications. Therefore, we consider four models with service interruptions as examples. In particular, in section 1.7, a discrete-time heterogeneous multiserver queueing system with a regenerative input flow and service interruptions, which are described by independent renewal processes for various servers, is discussed. It is shown that the traffic rate ρ is not expressed in terms of the first moments of the random variables defining the model for the preemptive repeat different service discipline (Gaver 1962). We also prove that $Q(t) \to \infty$ as $t \to \infty$ when $\rho \geq 1$ and $Q(t)$ is a stochastically bounded process when $\rho < 1$.

We should consider that models with service interruptions occur in numerous applications including manufacturing processes, multiprocessor computer networks, telecommunicating networks and various types of service counters.

White and Christie (1958) were the first to study queueing systems with interruptions. Those authors investigated the $M|M|1$ model with preemptive resume priority discipline. Their results were later extended by (Avi-Itzhak and Naor 1963) and (Thiruvengadam 1963) who study queues with general service times. Gaver (1962) studied single-server queues with batch Poisson arrivals and generally distributed service times. So far there is extensive literature concerning queueing systems with interruptions. There are some review papers that cover most of the

literature in this sphere. Some of the important papers on the single-server case are presented in Fiems and Bruneel (2013).

The most extensive literature survey on systems with interruptions both for single-server and multiserver cases is given by (Krishnamoorthy *et al.* 2012). This paper also covers some non-Markovian multichannel systems with homogeneous servers. There are some other articles with extensive literature survey as well (Pechinkin *et al.* 2009; Morozov *et al.* 2011). Nevertheless, to the best of our knowledge, there are no papers that study the stability problem for multichannel queueing systems with heterogeneous servers in non-Markovian case with general input flow and service times. Synchronization method combined with the regenerative theory is one of the powerful approaches to obtain stability conditions for such systems.

Let us also mention the fluid approximation approach as an alternative to the synchronization approach followed here. Such an approach has lent to significant progress in stability analysis of multiclass queueing networks (Dai 1995; Chen 1995; Chen and Yao 2001). See also Foss and Konstantopoulos (2004) for a survey of various approaches to stability of queueing systems with a focus on the fluid approach. Nevertheless, our analysis does not rely on a fluid approach because to the best of our knowledge, the synchronization method with regard to regenerative structure of the processes turns out to be suitable for obtaining complete and transparent proofs as well as natural stability conditions for the model at hand.

We also note that stability analysis is an essential and challenging stage of investigation of a stochastic model, however, stability conditions may be of independent interest. In particular, the stability criterion of the multiserver model can be used for the capacity planning of the model at the design stage to obtain the lower bound of the capacity that keeps system stable. As an example of applications of our results, we estimate the carrying capacity of the automobile road, intersected by a crosswalk. Under capacity we mean the upper bound of the intensity of the flow of cars when the queue of cars does not tend to infinity. This means that the analysis can be based on the results obtained in this chapter.

1.2. Model description

We begin from the definition of the regenerative flow (Afanasyeva and Bashtova 2014)). Assume a stochastic process $\{X(t), t \geq 0\}\,(X(0) = 0)$ taking values $0, 1, 2\ldots$ to be defined in a probability space $(\Omega, \mathscr{F}, P)$. The process has non-decreasing and right-continuous sample paths. There exists filtration $\{\mathscr{F}^X_{\leq t}\}_{t\geq 0}$, $(\mathscr{F}^X_{\leq t} \in \mathscr{F}$ for all $t \geq 0)$ exists such that X is measurable with respect to $\{\mathscr{F}^X_{\leq t}\}_{t\geq 0}$.

DEFINITION 1.1.– *The stochastic flow X is called regenerative if there is an increasing sequence of Markov moments $\{\theta_j^{(1)}, j \geq 0\}(\theta_0^{(1)} = 0)$ with respect to $\{\mathscr{F}_{\leq t}^X, t \geq 0\}$ such that the sequence*

$$\{\varkappa_j\}_{j=1}^{\infty} = \{X(\theta_{j-1}^{(1)} + t) - X(\theta_{j-1}^{(1)}), \theta_j^{(1)} - \theta_{j-1}^{(1)}, t \in (0, \theta_j^{(1)}) - \theta_{j-1}^{(1)}]\}_{j=1}^{\infty}$$

consists of independent identically distributed (iid) random elements on $(\Omega, \mathscr{F}, P)$.

The random variable $\theta_i^{(1)}$ is said to be the ith regeneration point of X and $\tau_i = \theta_i^{(1)} - \theta_{i-1}^{(1)}$ is the ith regeneration period $(i = 1, 2, \ldots)$. Let $\xi_j^{(1)} = X(\theta_j^{(1)}) - X(\theta_{j-1}^{(1)})$ be the number of customers arrived during the jth regeneration period. Assume that $\mathbf{E}\tau_1^{(1)} < \infty, \mathbf{E}\xi_1^{(1)} < \infty$. The limit $\lambda_X = \lim\limits_{t\to\infty} \frac{X(t)}{t}$ with probability 1 (w.p.1) is called the rate of X. It is easy to prove that $\lambda_X = \frac{\mathbf{E}\xi_1}{\mathbf{E}\tau_1^{(1)}}$ (Smith 1955; Thorisson 2000). The class of regenerative flows contains most of fundamental flows that are exploited in the queueing theory. First of all, the doubly stochastic Poisson process (Grandell 1976) with a regenerative process as a stochastic intensity is a regenerative flow. There are many other examples of regenerative flows, for instance, semi-Markovian, Markov-arrival, Markov-modulated and other processes (Afanasyeva and Tkachenko 2012).

We consider discrete-time queueing systems as well as continuous-time queueing systems (Avi-Itzhak and Naor 1963). In the first case, time is divided into fixed length intervals or slots and all arrivals and departures are synchronized with respect to slot boundaries. Moreover, in the case of some events being synchronized at one slot these events are ordered as follows: arrival and departure. The system is observed at the end of the slot.

We consider a queueing system S with m servers, FCFS discipline and regenerative input flow X. We assume that a server may simultaneously serve only one customer so that at any time the number of customers on the servers is not more than m. A customer leaves a system only after completion the service. The system is defined by the sequence $\{\overrightarrow{\eta}_n\}_{n=1}^{\infty}$ consisting of iid random vectors and multidimensional stochastic process $\{V(t), t \geq 0\}$, which are independent of the input flow X. The vector $\overrightarrow{\eta}_n$ determines the characteristics of the nth customer, that is service times by various servers or necessary number of servers for service. The process V describes the states of the servers. For example, for the systems with unreliable servers this process defines the moments of breakdowns and restorations of the servers. The state of the system at time t is described by the stochastic process $\overrightarrow{Z}(t) = (Z_1(t), \ldots, Z_k(t))$ where one of the coordinates is the number of customers in the system. We assume that the relation

$$\overrightarrow{Z}(t) = \Phi(\overrightarrow{Z}(0), t, \{X(s), s \leq t\}, \{\overrightarrow{\eta}_n\}_{n=1}^{X(t)}, \{V(s), s \leq t\}) \qquad [1.1]$$

takes place for some function $\Phi(\cdot)$ on the corresponding space. For example, for the system $Reg|G|1|\infty$ we consider the process $\overrightarrow{Z}(t) = (W(t), Q(t))$ where $W(t)$ is the virtual waiting time and $Q(t)$ is the number of customers in the system at time t. Let $\{\eta_j\}_{j=1}^{\infty}$ and $\{t_j\}_{j=1}^{\infty}$ be the sequences of service and arrival times of customers, respectively, and $U(t) = \sum_{j=1}^{X(t)} \eta_j - t$. Assuming that $Q(0) = 0$, we have $W(t) = \sup_{0 \le s \le t} (U(t) - U(s))$, $Q(t) = \sum_{j=1}^{X(t)} \mathbb{I}(t_j + W(t_j) + \eta_j > t)$ where $\mathbb{I}(\cdot)$ is an indicator function (Borovkov 1976).

The main goal of this chapter is the determination of the conditions of the stochastic boundedness of the number of customers $Q(t)$ in the system as $t \to \infty$. Our analysis is based on the construction of the auxiliary service process.

1.3. Auxiliary service process

For the system S, we define an auxiliary system S_0 with input flow X_0 such that when the number of customers in the system becomes less than m a new customer immediately arrives in the system. Therefore, there are always customers for service in S_0. Other characteristics such as the initial state, the sequence $\{\overrightarrow{\eta}_n\}_{n=1}^{\infty}$, stochastic process $\{\overrightarrow{V}(t), t \ge 0\}$ and a functional Φ are the same as for the system S. If in the system S the initial number of customers $Q(0) < m$, then the process X_0 has the jump $m - Q(0)$ at zero instant. We determine an auxiliary service process $Y(t)$ as the number of customers served in S_0 during $(0, t)$. Since the flow Y is defined by the processes $\{\overrightarrow{\eta}_n\}_{n=1}^{\infty}$ and V and these processes do not depend on the input flow X at the system S, we conclude that X and Y are independent flows.

We also need additional assumptions.

CONDITION 1.1.– For the continuous-time case, Y is a strongly regenerative flow with the sequence $\{\theta_n^{(2)}\}_{n=0}^{\infty} (\theta_0^{(2)} = 0)$ as points of regeneration.

We call the regenerative flow Y strongly regenerative if the regeneration period $\tau_n^{(2)} = \theta_n^{(2)} - \theta_{n-1}^{(2)}$ has the form

$$\tau_n^{(2)} = v_n^{(1)} + v_n^{(2)} \qquad [1.2]$$

where $P(v_n^{(1)} > x) = e^{-\delta x} (\delta \in (0, \infty))$, $v_n^{(1)}$ and $v_n^{(2)}$ are independent random variables and $Y(\theta_{n-1}^{(2)} + v_n^{(1)}) = Y(\theta_{n-1}^{(2)})$.

CONDITION 1.2.– For the discrete-time case, processes X and Y are regenerative aperiodic flows. As usually, aperiodicity means that the greatest common divisor (GCD)

$$GCD\{k : P(\theta_1^{(i)} = k) > 0\} = 1, \quad i = 1, 2.$$

Then we may determine common points of regeneration $\{T_n\}_{n=1}^{\infty}$ for both processes X and Y letting in the discrete-time case

$$T_n = \min\left\{\theta_j^{(1)} > T_{n-1} : \bigcup_{l=1}^{\infty}\{\theta_j^{(1)} = \theta_l^{(2)}\}\right\}, T_0 = 0 \qquad [1.3]$$

and in the continuous-time case

$$T_n = \min\left\{\theta_j^{(1)} > T_{n-1} : \bigcup_{l=1}^{\infty}\{\theta_{l-1}^{(2)} < \theta_j^{(1)} \leq \theta_{l-1}^{(2)} + v_l^{(1)}\}\right\}, T_0 = 0. \qquad [1.4]$$

LEMMA 1.1.– Let for the continuous-time (discrete-time) condition 1.1 (condition 1.2) be fulfilled. Then the sequence $\{T_n\}_{n=0}^{\infty}$ consists of common regeneration points for X and Y and

$$\mathbf{E}T_1 = \delta\mathbf{E}\tau_1^{(1)}\mathbf{E}\tau_1^{(2)} < \infty \qquad [1.5]$$

for the continuous-time case,

$$\mathbf{E}T_1 = \mathbf{E}\theta_1^{(1)}\mathbf{E}\theta_1^{(2)} < \infty \qquad [1.6]$$

for the discrete-time case.

PROOF.– Since the proof of [1.5] is almost the same as the proof of [1.6], we consider the discrete-time case only. Let

$$\nu_k = \min\left\{j > \nu_{k-1} : \bigcup_{l=1}^{\infty}\{\theta_j^{(1)} = \theta_l^{(2)}\}\right\}, \nu_0 = 0,$$

so that $T_k = \theta_{\nu_k}^{(1)}$. Then $\{\nu_k - \nu_{k-1}\}_{k=1}^{\infty}$ is a sequence of iid random variables and in accordance with Wald's identity $\mathbf{E}T_1 = \mathbf{E}\theta_1^{(1)}\mathbf{E}\nu_1$ (Feller 1971). Therefore, we need

to prove the finiteness of $\mathbf{E}\nu_1$. Denote by $h_2(t)$ $(h(t))$ the mean number of renewals at time t for the renewal process $\{\theta_n^{(2)}\}_{n=0}^{\infty}$ $(\{\nu_k\}_{k=0}^{\infty})$ so that

$$h_2(t) = \sum_{l=0}^{\infty} P(\theta_l^{(2)} = t)$$

and

$$h(t) = \sum_{k=0}^{\infty} P(\nu_k = t).$$

Taking into account condition 1.2, we derive from Blackwell's theorem (Thorisson 2000)

$$h_2(t) \xrightarrow[t\to\infty]{} \frac{1}{\mathbf{E}\theta_1^{(2)}}, \quad h(t) \xrightarrow[t\to\infty]{} \frac{1}{\mathbf{E}\nu_1}$$

if $\mathbf{E}\nu_1 < \infty$ and $h(t) \to 0$ as $t \to \infty$ if $\mathbf{E}\nu_1 = \infty$.

Because of X and Y independence

$$h(j) = P(\bigcup_{l=0}^{\infty}\{\theta_j^{(1)} = \theta_l^{(2)}\}) = \mathbf{E}\left(\sum_{l=0}^{\infty} P(\theta_j^{(1)} = \theta_l^{(2)} | \theta_j^{(1)})\right) = \mathbf{E}h_2(\theta_j^{(1)}).$$

Since $\theta_j^{(1)} \xrightarrow[j\to\infty]{} \infty$ w.p.1, then $h_2(\theta_j^{(1)}) \xrightarrow[j\to\infty]{} \frac{1}{\mathbf{E}\theta_1^{(2)}}$ w.p.1. Therefore, from Lebesgue's dominated convergence theorem, we obtain $\mathbf{E}\nu_1 = \mathbf{E}\theta_1^{(2)} < \infty$. ■

Later we consider both cases (discrete-time and continuous-time) together. We only have to take condition 1.2 instead of condition 1.1.

Let

$$\Delta_Y(n) = Y(T_n) - Y(T_{n-1}),$$
$$\Delta_X(n) = X(T_n) - X(T_{n-1}).$$

Then

$$\lambda_X = \frac{E\Delta_X(n)}{E(T_n - T_{n-1})}, \lambda_Y = \frac{E\Delta_Y(n)}{E(T_n - T_{n-1})}.$$

We define the traffic rate as follows:

$$\rho = \frac{\lambda_X}{\lambda_Y} = \frac{E\Delta_X(1)}{E\Delta_Y(1)}. \quad [1.7]$$

We think of λ_X and λ_Y as the arrival and service rate, respectively. Intuitively, it is clear that the number of customers in the system S is a stochastically bounded process if $\rho < 1$ and it is not the case if $\rho \geq 1$. The main stability result of this chapter consists of the formal proof of this fact.

We define the stochastic flow $\tilde{Y}(t)$ as the number of customers served at the system S during time interval $[0, t)$.

CONDITION 1.3.– The following stochastic inequalities take place:

$$\Delta_{\tilde{Y}}(n) = \tilde{Y}(T_n) - \tilde{Y}(T_{n-1}) \leq \Delta_Y(n),\ (n = 1, 2, \ldots).$$

Let $Q(t)$ be the number of customers in the system S including the customers on the servers at time t so that

$$Q(t) = Q(0) + X(t) - \tilde{Y}(t).$$

CONDITION 1.4.– There are two possible cases:

i) $Q(t)$ is a stochastically bounded process, i.e.

$$\lim_{x\to\infty} \liminf_{t\to\infty} P(Q(t) \leq x) = 1;$$

ii) $Q(t) \xrightarrow[t\to\infty]{P} \infty$.

CONDITION 1.5.– If $Q(t) \xrightarrow[t\to\infty]{P} \infty$, then for any $\epsilon > 0$ there exists n_ϵ such that for $n > n_\epsilon$

$$E\Delta_{\tilde{Y}}(n) \geq E\Delta_Y(1) - \epsilon. \quad [1.8]$$

1.4. Instability result for the case $\rho \geq 1$

THEOREM 1.1.– Let conditions 1.3 and 1.1 (1.2) for the continuous-time (for the discrete-time) case be fulfilled. If $\rho \geq 1$, then

$$Q(t) \xrightarrow[t\to\infty]{P} \infty. \quad [1.9]$$

PROOF.– Consider the embedded process $Q_n = Q(T_n)$ and denote $Z_k = \sum_{j=1}^{k} (\Delta_X(j) - \Delta_Y(j))$ $(Z_0 = 0)$.

Define the auxiliary sequence $\{\hat{Q}_k\}_{k=0}^{\infty}$ by the recursion

$$\hat{Q}_k = \max[0, \hat{Q}_{k-1} + \Delta_X(k) - \Delta_Y(k)],\ \hat{Q}_0 = Q(0).$$

Because of the equality $Q_k = Q_{k-1} + \Delta_X(k) - \Delta_{\tilde{Y}}(k)$ and condition 1.3 we get the stochastic inequality $Q_k \geq \hat{Q}_k$. It is well known (Feller 1971) that

$$\hat{Q}_k = \max_{0 \leq j \leq k} (Q(0) + Z_k, Z_k - Z_j)$$

and in distribution

$$\hat{Q}_k \geq \max_{0 \leq j \leq k} Z_j.$$

For $\rho \geq 1$, the sequence $\{Z_j\}_{j \geq 0}$ is a random walk with a non-negative drift. Hence, except when $\Delta_X(1) = \Delta_Y(1) = c$ (c is a constant), $\max_{0 \leq j \leq k} Z_j \xrightarrow[k \to \infty]{} \infty$ (w.p.1) (Feller 1971) that completes the proof. ■

1.5. Stochastic boundedness for the case $\rho < 1$

Our objective here is to establish stochastic boundedness of the process Q when the traffic rate $\rho < 1$. Under some additional assumptions providing the regenerative structure of the process Q, this property has its stability as a consequence.

THEOREM 1.2.– Let conditions 1.4 and 1.5 and condition 1.1 (1.2) for the continuous-time (for the discrete-time) case be fulfilled. If $\rho < 1$, then Q is a stochastically bounded process.

PROOF.– Because of condition 1.4 there are two possible cases: $Q_n = Q(T_n)$ is either stochastically bounded or $Q_n \xrightarrow[n \to \infty]{P} \infty$. Assume that the second case takes place and $\rho < 1$. Because of condition 1.5 for $0 < \epsilon < E(\Delta_Y(1) - \Delta_X(1))$, there is n_ϵ such that for $n > n_\epsilon$

$$EQ_{n+1} = EQ_n + E\Delta_X(1) - E\Delta_{\tilde{Y}}(n) \leq$$
$$\leq EQ_n + E\Delta_X(1) - E\Delta_Y(1) + \epsilon \leq EQ_n$$

that contradicts our assumption that $Q_n \xrightarrow[n \to \infty]{P} \infty$. ■

In the next sections, we discuss some examples to verify our results and to compare them with previous works. First, we consider two queueing models with service interruptions. These models occur in numerous applications and there is extensive literature concerning queueing system with interruptions. Let us mention some papers in which detail description of the literature in this sphere is given (Krishnamoorthy *et al.* 2012; Fiems and Bruneel 2013; Pechinkin *et al.* 2009; Morozov *et al.* 2011). However, to the best of our knowledge there are no papers that study the stability problem for multichannel systems with heterogeneous servers for the non-Markovian case: with general input flow and general distribution of blocked and available periods.

1.6. Queueing system with unreliable servers and preemptive resume service discipline

We consider a continuous-time queueing system with regenerative input flow X and m heterogeneous servers that may be not available for operation from time to time. We also propose that the velocity of the service may be dependent on the state of the server. Assume that for the ith server a stochastic process $n_i(t)$ with state space $(0; r_1^{(i)}; \ldots; r_{k_i}^{(i)})$, $r_j^{(i)} > 0$, $j = \overline{1, k_i}$ is defined. If $n_i(t) = 0$, then the ith server is in unavailable state, for instance it is broken; if $n_i(t) = r_j^{(i)}$, then the ith server is working with the velocity $r_j^{(i)}$ ($j = \overline{1, k_i}$, $i = \overline{1, m}$). Service times of customers by the ith server in the case when the velocity of the service is equal to one constitute a sequence $\{\eta_{in}\}_{n=1}^{\infty}$ of iid random variables, which does not depend on the input flow and service times by other servers, $b_i = E\eta_{in} < \infty$, $B_i(x) = P(\eta_{in} \le x)$, $i = \overline{1, m}$.

It is possible that an unavailable period starts while a customer is receiving service. Then service of the customer is immediately interrupted. There are various disciplines for continuation of the service after restoration (Gaver 1962). Here, we consider the preemptive resume service discipline assuming that interrupted service continues when the server returns from a blocked period and the service velocity is the next state of the process $n_i(t)$.

CONDITION 1.6.– The stochastic process $\overrightarrow{n}(t) = (n_1(t), \ldots, n_m(t))$ is strongly regenerative with regeneration points $\{\theta_n^{(2)}\}_{n=1}^{\infty}$ ($\theta_0^{(2)} = 0$), $\tau_n^{(2)} = \theta_n^{(2)} - \theta_{n-1}^{(2)}$, $E\tau_n^{(2)} < \infty$ with an exponential phase $v_n^{(1)}$ so that $\tau_n^{(2)} = v_n^{(1)} + v_n^{(2)}$. We also assume that $n_i(\theta_{n-1}^{(2)} + t) = 0$ for $t \in [0, v_n^{(1)}]$, $i = \overline{1, m}$.

It follows from condition 1.6 and Smith's (1955) theorem that there exist the limits

$$\lim_{t\to\infty} P(\overrightarrow{n}(t) = (j_1, \ldots, j_m)) = \pi_{j_1,\ldots,j_m}$$

and

$$\lim_{t\to\infty} P(n_i(t) = j_i) = \pi_{j_i}^{(i)},$$

where j_i takes values $0, r_1^{(i)}, \ldots, r_{k_i}^{(i)}$, $i = \overline{1,m}$.

To define an auxiliary process $Y_i(t)$ for the ith server, we introduce a counting process

$$K_i(t) = \max\{j : \sum_{l=1}^{j} \eta_{il} \le t\}.$$

Then

$$Y_i(t) = K_i(\int_0^t n_i(y)dy) \qquad [1.10]$$

and

$$Y(t) = \sum_{i=1}^{m} Y_i(t).$$

CONDITION 1.7.– Service times have the first exponential phase, i.e.

$$\eta_{in} = \eta_{in}^{(1)} + \eta_{in}^{(2)}$$

where $\eta_{in}^{(1)}$ and $\eta_{in}^{(2)}$ are independent random variables and $P(\eta_{in}^{(1)} \le x) = e^{-\alpha_i x}$ $(\alpha_i \in (0,\infty))$.

As regeneration points for Y we take subsequence $\{\theta_{n_k}^{(2)}\}_{k=1}^{\infty}$ of the sequence $\{\theta_n^{(2)}\}_{n=1}^{\infty}$ such that at time $\theta_{n_k}^{(2)}$ interrupted services for processes Y_i $(i = \overline{1,m})$ are in the exponential phase. Because of conditions 1.6 and 1.7, Y is a strongly regenerative flow and we may define the common sequence $\{T_n\}_{n=1}^{\infty}$ of regeneration points for both processes X and Y with the help of formula [1.4]. We need only to take $\{\theta_{n_k}^{(2)}\}_{k=1}^{\infty}$ instead of $\{\theta_n^{(2)}\}_{n=1}^{\infty}$.

Because of [1.10] we can easily obtain from the renewal theory the formula for the rate of the auxiliary process

$$\lambda_Y = \sum_{i=1}^{m} b_i^{-1} \sum_{j=1}^{k_i} r_j^{(i)} \pi_j^{(i)}. \qquad [1.11]$$

Now we may calculate the traffic rate ρ and under some assumptions we get the necessary and sufficient stability condition for the system based on theorems 1.1 and 1.2. As an example, we consider the famous case (Morozov *et al.* 2011) when $r_i^{(j)} = 1$, $j = \overline{1,k_i}$, $i = \overline{1,m}$, i.e. a server may be in an available or unavailable state. Let $\{s_{i,n}^{(2)}\}_{n=1}^{\infty}$ be moments of breakdowns and $\{s_{i,n}^{(1)}\}_{n=1}^{\infty}$ moments of restorations for the ith server. Here

$$0 = s_{i,0}^{(2)} < s_{i,1}^{(1)} < s_{i,1}^{(2)} < \dots . \qquad [1.12]$$

Then $u_{i,n}^{(1)} = s_{i,n}^{(1)} - s_{i,n-1}^{(2)}$ and $u_{i,n}^{(2)} = s_{i,n}^{(2)} - s_{i,n}^{(1)}$ denote the length of the nth blocked and the nth available period for the ith server, respectively, $(i = \overline{1,m})$. The sequence $\{u_{i,n}^{(1)}, u_{i,n}^{(2)}\}_{n=1}^{\infty}$ consists of iid random vectors (for all $(i = \overline{1,m})$) and these sequences do not depend on the input flow X and service times. Let $u_{i,n} = u_{i,n}^{(1)} + u_{i,n}^{(2)}$ be the length of the nth cycle for the server i. A cycle consists of a blocked period followed by an available period. We assume that

$$Eu_{i,n}^{(1)} = a_i^{(1)} < \infty,\ Eu_{i,n}^{(2)} = a_i^{(2)} < \infty,\ a_i = a_i^{(1)} + a_i^{(2)}\ (i = \overline{1,m}).$$

We put $n_i(t) = 0$ if the ith server is in an unavailable state at time t and $n_i(t) = 1$, otherwise $(i = \overline{1,m})$. If a blocked period $u_{i,n}^{(1)}$ has an exponential phase, i.e. $u_{i,n}^{(1)} = v_{i,n}^{(1)} + v_{i,n}^{(2)}$ where $v_{i,n}^{(1)}$ and $v_{i,n}^{(2)}$ are independent random variables and $v_{i,n}^{(1)}$ has an exponential distribution with a parameter α_i, then we may define the sequence $\{\theta_n^{(2)}\}_{n=1}^{\infty}$ of regeneration points for the regenerative process $\overrightarrow{n}(t) = (n_1(t), \dots, n_m(t))$ as above. Therefore, condition 1.6 holds. Under condition 1.7, the auxiliary process Y is strongly regenerative and we can construct the common points of regeneration $\{T_n\}_{n=1}^{\infty}$ for X and Y and apply theorems 1.1 and 1.2 for this model. Since

$$\pi_i = \lim_{t\to\infty} P(n_i(t) = 1) = \frac{a_i^{(2)}}{a_i}$$

we have from [1.11]

$$\rho = \frac{\lambda_X}{\sum\limits_{i=1}^{\infty} b_i^{-1} \frac{a_i^{(2)}}{a_i}}.$$

If $b_i = b$, then we get the same stability condition as obtained in Morozov *et al.* (2011) for a queueing system $GI|G|m$ with a common distribution function of service times for all servers.

COROLLARY 1.1.– For a queueing system with $r_j^{(i)} = 1$, $j = \overline{1, k_i}$, $i = \overline{1, m}$

$$Q(t) \xrightarrow[t\to\infty]{P} \infty$$

if $\rho > 1$.

Under condition 1.4, the process is stochastically bounded if $\rho < 1$.

PROOF.– Let, as before, $\tilde{Y}_i(t)$ be the number of customers actually served on the ith server up to time t. It is evident that stochastic inequality

$$Y_i(t) \geq \tilde{Y}_i(t),\ i = \overline{1, m}$$

for $t \geq 0$ takes place and hence

$$Q(t) = Q(0) + X(t) - \tilde{Y}(t) \geq Q(0) + X(t) - Y(t).$$

Since $\rho > 1$, then $Q(t) \xrightarrow[t\to\infty]{P} \infty$.

To prove the second statement, we first assume that conditions 1.6 and 1.7 hold. Then condition 1.1 for the process Y takes place. We also may organize the performance of the systems S and S_0 in such a way that inequality [1.8] is realized when $Q(t) \xrightarrow[t\to\infty]{P} \infty$. Thus, conditions 1.1, 1.4 and 1.5 are satisfied and because of theorem 1.2 the process Q is stochastically bounded.

If conditions 1.6 and 1.7 (or one of them) are not valid, we construct a system S_δ satisfying conditions 1.6 and 1.7 and majorising our system S, so that in distribution

$$Q(t) \leq Q_\delta(t) + m. \qquad [1.13]$$

Here, $Q_\delta(t)$ is the number of customers in the system S_δ at instant t. Let us introduce independent sequences $\{\{v_{i,n}\}_{n\geq 1}, \{\gamma_{i,n}\}_{n\geq 1}\}_{i=1}^m$ of iid random variables with exponential distribution with a rate δ. Assume that repair time $\tilde{u}_{i,n}^{(1)}$ in the system S_δ has the form $\tilde{u}_{i,n}^{(1)} = v_{i,n}^{(1)} + u_{i,n}^{(1)}$ and service time $\tilde{\eta}_{i,n}$ by the ith server has the form $\tilde{\eta}_{i,n} = \eta_{i,n} + \gamma_{i,n}$.

Then S_δ satisfies conditions 1.6 and 1.7. Since $\rho_\delta = \lambda_X (\sum\limits_{i=1}^{m} \frac{\delta}{1+\delta b_i} \frac{a_i^{(2)}\delta}{1+\delta a_i})^{-1}$ and $\rho = \rho_\delta < 1$ we may choose δ so that $\rho_\delta < 1$.

The proof of [1.13] is based on the "so-called" probability space method (Belorusov 2012).

Let us note that condition 1.4 may be provided in various ways. For instance, assume that blocked (or available) period has an exponential phase and

$$B_1(x) > 0 \text{ for all } x > 0. \qquad [1.14]$$

Then Q is a regenerative process with points of regeneration $\{\theta_{n_k}^{(1)}\}_{k=1}^{\infty}$ that is a subsequence of the sequence $\{\theta_n^{(1)}\}_{n=1}^{\infty}$ such that $Q(\theta_{n_k}^{(1)}) = 0$ and all servers are in the exponential phase of their blocked (or available) periods. Now condition 1.4 follows directly from theorem 1 in Afanasyeva and Tkachenko (2014). We also note that in this case Q is a stable process if $\rho < 1$. If only assumption [1.14] takes place with the help of the majorising system S_δ, we obtain the stochastic boundedness Q when $\rho < 1$. ■

1.7. Discrete-time queueing system with interruptions and preemptive repeat different service discipline

Here, we consider the system with interruptions described in the previous section for the discrete-time case. The moments of breakdowns $\{s_{i,n}^{(2)}\}_{n=1}^{\infty}$ and moments of restorations $\{s_{i,n}^{(1)}\}_{n=1}^{\infty}$ for the ith server satisfy [1.12]. The input flow X is an aperiodic discrete-time regenerative flow with rate λ_X.

We consider the preemptive repeat different service discipline that means that the service is repeated from the start after restoration of the server and the new service time is independent of the original service time (Gaver 1962).

To define the process Y_i for the ith server in the auxiliary system S_0, we introduce the collection $\left\{\{\eta_{i,n}^{(j)}\}_{n=1}^{\infty}\right\}_{j=1}^{\infty}$ of independent sequences $\{\eta_{i,n}^{(j)}\}_{n=1}^{\infty}$ consisting of iid random variables with distribution function B_i. Of course, we assume that $P(\eta_{i,n}^{(1)} > u_{i,n}^{(2)}) > 0$, $(i = \overline{1,m})$. Let $\mathscr{K}_{i,j}(t)$ be the counting process associated with the sequence $\{\eta_{i,n}^{(j)}\}_{n=1}^{\infty}$, i.e. $\mathscr{K}_{i,j}(t) = \max\{k : \sum_{n=1}^{k} \eta_{i,n}^{(j)} \le t\}$, $(\mathscr{K}_{i,j}(0) = 0)$ and $\mu_i(t)$ be the number of cycles for the ith server during $[0,t]$, i.e. $\mu_i(t) = \max\{j : \sum_{n=1}^{j} u_{i,n} \le t\}$, $(\mu_i(0) = 0)$. Then the process Y_i is defined by the relation

$$Y_i(t) = \sum_{j=1}^{\mu_i(t)} \mathscr{K}_{i,j}(u_{i,j}^{(2)}) + \mathscr{K}_{i,\mu_i(t)+1}(\max[0, t - s_{i,\mu_i(t)+1}^{(1)}]) \qquad [1.15]$$

and $Y(t) = \sum_{i=1}^{m} Y_i(t)$. We denote by $H_i(t)$ the renewal function for $\mathscr{K}_{i,j}$, i.e. $H_i(t) = \mathbf{E}\mathscr{K}_{i,j}(t)$.

LEMMA 1.2.– There exists the limit

$$\lambda_{Y_i} = \lim_{t\to\infty} \frac{Y_i(t)}{t} = \frac{\mathbf{E}H_i(u_{i,n}^{(2)})}{a_i} \quad \text{w.p.1.}$$

The proof easily follows from the evident inequalities

$$g_i(\mu_i(t)) \le Y_i(t) \le g_i(\mu_i(t)+1)$$

where $g_i(n) = \sum_{j=1}^{n} \mathscr{K}_{i,j}(u_{i,j}^{(2)})$, the strong law of large numbers and convergence $t^{-1}\mu_i(t) \xrightarrow[t\to\infty]{} a_i^{-1}$ w.p.1.

From lemma 1.2, we have

$$\lambda_Y = \lim_{t\to\infty} \frac{Y(t)}{t} = \sum_{i=1}^{m} \frac{\mathbf{E}H_i(u_{i,1}^{(2)})}{a_i}. \qquad [1.16]$$

We introduce the counting processes

$$N_0(t) = \max\{k : \theta_k^{(1)} \le t\},$$

$$N_i(t) = \max\{k : s_{i,k}^{(2)} \le t\}, \quad i = \overline{1,m}.$$

CONDITION 1.8.– The counting processes $N_0(t)$ and $N_i(t)$ ($i = \overline{1,m}$) are aperiodic.

Then Y is a regenerative aperiodic flow with points of regeneration

$$\theta_j^{(2)} = \min\{t > \theta_{j-1}^{(2)} : \bigcap_{i=1}^{m} [N_i(t) - N_i(t-1) > 0]\}, \quad \theta_0^{(2)} = 0.$$

In other words, $\theta_j^{(2)}$ is a point of regeneration of Y if all the servers get out of the order simultaneously at this moment. Taking into account condition 1.8, we conclude from lemma 1.1 that $\mathbf{E}(\theta_j^{(2)} - \theta_{j-1}^{(2)}) < \infty$. Now we construct the sequence $\{T_n\}_{n=1}^{\infty}$ of common points of regeneration for processes X and Y with the help of [1.3]. Because of lemma 1.1 $\mathbf{E}(T_n - T_{n-1}) < \infty$ and the traffic rate ρ of the system is defined by [1.7].

COROLLARY 1.2.– Let condition 1.8 be fulfilled. Then

i) $Q(t) \xrightarrow[t\to\infty]{P} \infty$ if $\rho \geq 1$;

ii) $Q(t)$ is a stochastically bounded process if $\rho < 1$.

PROOF.– The first statement follows from theorem 1.1 since conditions 1.2 and 1.3 are realized.

Let $\rho < 1$. For the system S, we introduce the embedded process $x_n = (Q_n, \zeta_1(n), \ldots, \zeta_m(n))$, $n \geq 0$, where Q_n is the number of customers in the system on time T_n and $\zeta_i(n) = 1$ if there is a customer on the ith server and $\zeta_i(n) = 0$ otherwise. In a view of the service discipline after service restoration and properties of the synchronization epochs $\{T_n\}_{n\geq 0}$, the process $\{x_n\}_{n\geq 1}$ is a Markov chain with countable set of states $\mathscr{R} = \{\{0\}, (j, e_1, \ldots, e_m),\ j = \overline{1, m-1}; \{j\},\ j \geq m\}$. Let R_0 be the set of unessential states and R_j $(j = \overline{1, r})$ irreducible classes of communicating states. It follows from the condition $\rho < 1$ that the number of classes $r < \infty$.

For any aperiodic class $\mathscr{K}_l$ of states based on Foster's criterion (Meyn and Tweedie 2009), we may easily prove that this class is ergodic (Afanasyeva and Tkachenko 2016, 2018). Therefore, the process Q_n is stochastically bounded if $Q_0 \in \mathscr{K}_l$. It is also true if $\mathscr{K}_l$ is a periodic class. Since the number of classes $r < \infty$, we obtain the stochastic boundedness of the process Q_n and therefore $Q(t)$.

We may obtain the upper bound of the traffic rate ρ providing the stochastic boundedness of the process Q. It is known from (Borovkov 1976) that

$$H_i(t) \geq \frac{t}{b_i} - 1.$$

Therefore

$$\sum_{i=1}^{m} H_i(u_{i,n}^{(2)}) \geq \sum_{i=1}^{m} \frac{a_i^{(2)}}{b_i a_i} - \sum_{i=1}^{m} \frac{1}{a_i}$$

and sufficient condition of the stochastic boundedness of Q has the following form

$$\lambda + \sum_{i=1}^{m} a_i^{-1} \leq \sum_{i=1}^{m} \frac{a_i^{(2)}}{b_i a_i}.$$

If $b_i = b$, then we have the same condition as obtained in Morozov *et al.* (2011).

■

1.8. Queueing system with a preemptive priority discipline

In this section we study a continuous-time queueing system with two independent regenerative input flows X_1 and X_2 with intensities λ_1 and λ_2 and m servers. The customers of the second type (which belong to X_2) have an absolute priority with respect to customers of the first type. Service interruption for the low priority customer occurs when a high priority customer arrives during a low priority customer's service time. If at an arrival time of the second type customer there are m_1 free servers, m_2 servers occupied by customers of the first type and $m - m_1 - m_2$ servers occupied by customers of the second type, then an arriving customer randomly chooses any server from $m_1 + m_2$ servers, which are not busy by customers of the second type. Service times by the ith server for high(low) priority customers have distribution function B_0 (B_i, $i = \overline{1,m}$) with mean b_0 (b_i, $i = \overline{1,m}$). Therefore, for high priority customers we have a system $Reg|G|m$ with homogeneous servers and for low priority customers a system with interruptions and preemptive resume service discipline considered in section 1.6.

Denote by $Q_i(t)$ the number of customers of the ith type at the system including the customers on the servers at time t ($i = 1, 2$). Let $\{\theta_j^{(1)}\}_{j=1}^{\infty}$ ($\theta_0^{(1)} = 0$) and $\{\theta_j^{(2)}\}_{j=1}^{\infty}$ ($\theta_0^{(2)} = 0$) be the sequences of regeneration points for X_1 and X_2, respectively. Under some additional conditions, for example, when the inequality [1.14] is valid for the function B_0 (other sufficient assumptions are given in Afanasyeva and Tkachenko (2014)), the process Q_2 is regenerative with points of regeneration

$$S_n^{(2)} = \min\{\theta_j^{(2)} > S_{n-1}^{(2)} : Q_2(\theta_j^{(2)}) = 0\},\ n > 0;$$

$$S_0^{(2)} = 0.$$

The stability condition for the process Q_2 has the form (Afanasyeva and Tkachenko 2014)

$$\rho_2 = \frac{\lambda_2 b_0}{m} < 1 \qquad [1.17]$$

that is supposed to be fulfilled. We now want to get the stability condition for the process Q_1.

We start with the definition of the process of interruptions. Let $n_i(t) = 0$ if at instant t the ith server is occupied by a high priority customer and $n_i(t) = 1$ otherwise, $i = \overline{1,m}$. As regeneration points for $\overrightarrow{n}(t) = (n_1(t), \ldots, n_m(t))$, we take

subsequence $\{\theta_{j_k}^{(2)}\}_{k=1}^{\infty}$ of the regeneration points sequence $\{\theta_j^{(2)}\}_{j=1}^{\infty}$ for the input flow X_2 such that $Q_2(\theta_{j_k}^{(2)}) = 0$. As before, we assume that [1.14] holds for B_0. Since X_2 is a strongly regenerative flow, condition 1.6 is fulfilled.

To obtain the traffic rate for low priority customers, we need to find $\pi_i = \lim\limits_{t\to\infty} P(n_i(t) = 1)$. Because of the rule of the server choose by an arriving high priority customer, we have $\pi_i = \pi_1 = \pi$ for all $i = \overline{1,m}$. To calculate π, we define for high priority customers the following processes. Let $w_i(t)$ be the residual service time (virtual waiting time) on the ith server at instant t and $Z_i(t)$ the total service time of customers which arrived up to time t and have to be served on the ith server. Thus

$$\sum_{i=1}^{m} Z_i(t) = \sum_{j=1}^{X_2(t)} \eta_j$$

where η_j is the service time of the jth arrived customer. We note that w.p.1

$$\lim_{t\to\infty} \frac{Z_i(t)}{t} = \frac{1}{m} \lim_{t\to\infty} \sum_{j=1}^{X_2(t)} \eta_j = \frac{\lambda_2 b_0}{m}$$

and

$$\frac{w_i(t)}{t} \xrightarrow[t\to\infty]{} 0$$

because of the stability condition [1.17].

Since

$$w_i(t) = Z_i(t) - \int_0^t (1 - n_i(y))dy$$

and w.p.1

$$\lim_{t\to\infty} t^{-1} \int_0^t (1 - n_i(y))dy = 1 - \pi$$

then $\pi = 1 - \frac{\lambda_2 b_0}{m}$. Therefore, the traffic rate for low priority customers has the form

$$\rho_1 = \lambda_1((1 - \frac{\lambda_2 b_0}{m}) \sum_{i=1}^{m} b_i^{-1})^{-1}.$$

From corollary 1.1 we obtain corollary 1.3.

COROLLARY 1.3.– Assume that X_2 is a strongly regenerative flow, B_0 satisfies [1.14] and $\rho_2 = \frac{\lambda_2 b_0}{m} < 1$. Then $Q_1(t) \xrightarrow[t\to\infty]{P} \infty$ if $\rho_1 > 1$. If additionally B_i satisfies [1.14] and $\rho_1 < 1$, then $Q_1(t)$ is a stable process.

The first statement follows from corollary 1.1 since $\overrightarrow{n}$ is a strongly regenerative process. To prove the second statement, we note that Q_2 is a regenerative process and its stochastic boundedness means stability.

1.9. Queueing system with simultaneous service of a customer by a random number of servers

Here, we consider a system S assuming that the nth customer requires service from ζ_n server simultaneously ($\zeta_n = \overline{1,m}$). The customer arrived in an empty queue begins service immediately if the number of available servers is more or equal ζ_n. Otherwise the customer becomes the first in a queue and the service begins when the required number of servers becomes available. A customer who arrives in an nonempty queue takes the last place in the queue. When service begins, each server's completion time is independent of all other servers and has an exponential distribution with rate μ. The sequence $\{\zeta_n\}_{n=1}^{\infty}$ consists of iid random variables and $\alpha_j = P(\zeta = j)$, $j = \overline{1,m}$, $\sum\limits_{j=1}^{m} \alpha_j = 1$.

Queueing systems with simultaneous service have been studied in a number of works (Rumyantsev and Morozov 2017; and references therein). The stability conditions in an explicit form have been obtained for systems with a Poisson input flow and independent exponentially distributed service times by (Gillent and Latouche 1983). The main goal of this section is an extension of the stability condition to the model with a regenerative input flow X based on theorems 1.1 and 1.2. Thus, we consider the system S described in section 1.2 with $\overrightarrow{\eta}_n = (\eta_{n1}, \ldots, \eta_{n\zeta_n}, \zeta_n)$ where $\{\eta_{nj}\}_{j=1}^{m}$ is a sequence of independent exponentially distributed random variables not depending on ζ_n. Let S_0 be an auxiliary system defined in section 1.3. Instead of the process Y we consider the auxiliary flow Z that is the number of service completions by all m servers up to time t in the auxiliary system S_0. Denote by $U(t)$ the number of occupied servers at time t in the system S_0. Then U is a Markov chain and Z is a doubly stochastic Poisson process (Grandell 1976) with a random intensity $\lambda_Z(t) = \mu U(t)$. We note that the process U hits the state $\{m\}$ from any state $j = 1, 2, \ldots, m$ with a positive probability. It means that all states attainable from the state $\{m\}$ constitute the finite class $\mathcal{K}$ of communicating states. Therefore, there are limits

$$\lim_{t\to\infty} P(U(t) = j) = P_j > 0 \text{ for } j \in \mathcal{K}.$$

Let $s_0 = \max\{s = \overline{1,m} : \alpha_s > 0\}$ and $j_0 = m - s_0 + 1, \mathcal{K} = (j_0, j_0+1, \ldots, m)$.

We have the system of equations for $P_j = \lim\limits_{t\to\infty} P(U(t) = j)$, $j = \overline{j_0, m}$.

$$\begin{cases} jP_j = (j+1)\frac{\beta_j}{\beta_{j+1}}P_{j+1} \quad \text{for } j_0 \le j < m \\ m(1-\alpha_1)P_m = \sum\limits_{j=j_0}^{m-1} \frac{\alpha_{m-j+1}}{\beta_j} P_j \\ \sum\limits_{j=j_0}^{m} P_j = 1 \\ P_j = 0 \quad \text{for } j < j_0 \end{cases} \qquad [1.18]$$

where $\beta_j = \alpha_m + \alpha_{m-1} + \ldots + \alpha_{m-j+1}$, $j = \overline{1,m}$. We may easily verify that the solution of [1.18] has the form

$$P_j = \frac{C}{j} P(\zeta > m - j),\ C = (\sum_{j=1}^{m} \frac{1}{j}\beta_j)^{-1},\ j = \overline{1,m}.$$

Since

$$\lambda_Z = \lim_{t\to\infty} \frac{Z(t)}{t} = \mu \sum_{j=1}^{m} jP_j$$

we get

$$\lambda_Z = \frac{\mu E\zeta}{\sum\limits_{j=1}^{m} \frac{1}{j} P(\zeta > m - j)}$$

and the traffic rate for the system S

$$\rho = \frac{\lambda_X E\zeta}{\lambda_Z} = \frac{\lambda_X}{\mu} \sum_{j=1}^{m} \frac{1}{j} P(\zeta > m - j). \qquad [1.19]$$

Let us note that this formula is the same as obtained by (Gillent and Latouche 1983) for queueing systems with a Poisson input flow. To employ theorems 1.1 and 1.2, we have to verify conditions 1.1, 1.3, 1.4 and 1.5. First of all we note that Z is a strongly regenerative flow. As points of regeneration, we may take the sequential hitting times $\{t_n^{(j)}\}_{n=1}^{\infty}$ of $U(t)$ into the fixed state $j \in \mathcal{K}$. (We take $t_1^{(j)} = 0$ if

$U(0) = j$.) Then $t_{n+1}^{(j)} - t_n^{(j)} = v_n^{(j)} + \gamma_n^{(j)}$, where $v_n^{(j)}$ is the sojourn time in the state j and $\gamma_n^{(j)}$ is the return time to this state after exit from it for $U(t)$. The random variables $v_n^{(j)}$ and $\gamma_n^{(j)}$ are independent and $v_n^{(j)}$ has an exponential distribution. Moreover, $Z(t_{n+1}^{(j)} + v_n^{(j)}) = Z(t_{n+1}^{(j)})$, therefore, condition 1.1 holds. Let $q(t)$ be the total number of servers that are already busy or will be busy by service of the $Q(t)$ customers, which are present at the system S at time t. Then $q(t)$ as well as $Q(t)$ is a regenerative process with a sequence $\{\theta_{n_k}^{(1)}\}_{k=1}^{\infty}$ of points of regeneration that is a subsequence of $\{\theta_n^{(1)}\}_{n=1}^{\infty}$ such that $q(\theta_{n_k}^{(1)}) = Q(\theta_{n_k}^{(1)}) = 0$. Let us recall that $\{\theta_n^{(1)}\}_{n=1}^{\infty}$ is a sequence of points of regeneration for X. Therefore, because of theorem 1 from (Afanasyeva and Tkachenko 2014), condition 1.4 is fulfilled. Now for any fix $j \in \mathcal{K}$ we define the common points of regeneration $\{T_n^{(j)}\}_{n=1}^{\infty}$ for the input flow X and auxiliary flow Z by the relation

$$T_n^{(j)} = \min\{\theta_l^{(1)} > T_{n-1}^{(j)} : \bigcup_{s=1}^{\infty}\{t_s^{(j)} \le \theta_l^{(1)} < t_s^{(j)} + v_s^{(j)}\}\},\ T_0^{(j)} = 0.$$

Let

$$\Delta_Z^{(j)}(n) = Z(T_{n+1}^{(j)}) - Z(T_n^{(j)}),$$

$$\Delta_X^{(j)}(n) = X(T_{n+1}^{(j)}) - X(T_n^{(j)}), \qquad [1.20]$$

$$\Delta_{\tilde{Z}}^{(j)}(n) = \tilde{Z}(T_{n+1}^{(j)}) - \tilde{Z}(T_n^{(j)}),$$

where $\tilde{Z}$ is the process Z for the system S that is the number of service completions by all m servers up to time t. Now we formulate the main result of this section.

COROLLARY 1.4.– For the system S, the process Q is a stable process if and only if $\rho < 1$.

PROOF.– Consider the case $\rho \ge 1$ and take m instead of j in [1.20]. It is evident that the stochastic inequality

$$\Delta_{\tilde{Z}}^{(m)}(n) \le \Delta_Z^{(m)}(n) \quad (n = 1, 2, \ldots)$$

takes place. Therefore condition 1.3 is fulfilled. Based on theorem 1.1, we obtain the convergence

$$Q(t) \xrightarrow[t\to\infty]{P} \infty. \qquad [1.21]$$

For the case $\rho < 1$, we take j_0 instead of j in [1.20]. Then if convergence [1.21] takes place for any $\epsilon > 0$, there is n_ϵ such that

$$E\Delta_{\tilde{Z}}^{(j_0)}(n) \geq E\Delta_{Z}^{(j_0)}(1) - \epsilon \quad \text{for } n \geq n_\epsilon.$$

The proof is based on the approach described in Afanasyeva and Tkachenko (2014). Thus, conditions 1.1, 1.4 and 1.5 of theorem 1.2 are fulfilled and Q is a stochastically bounded process when $\rho < 1$. Since Q is a regenerative process, this means that it is stable.

We see that for the model under consideration, the stability condition does not depend on the structure of the input flow. ■

1.10. Applications to transport systems analysis

The study of traffic flows has a long history (Gideon and Pyke 1999; Grinbeerg 1959; Greenshields 1935; Inose and Hamada 1975; and references therein). Various methods such as cellular automate (Maerivoet and de Moor 2005), statistical mechanics and mathematical physics (Blank 2003; Chowdhury 1999; Fuks and Boccara 2001; Helbing 2001; Schadschneider 2000) or queueing theory (Afanasyeva and Bulinskaya 2009, 2010, 2011, 2013; Afanasyeva and Mihaylova 2015; Afanasyeva and Rudenko 2012; Baycal-Gursoy and Xiao 2004; Baycal-Gursoy *et al.* 2009; Caceres and Ferrari 2007) were used.

The purpose of the proposed study is an estimation of the carrying capacity of the automobile road, intersected by a crosswalk. Under the capacity, we mean the upper bound of the intensity of the flow of cars, when the queue of cars does not tend to infinity. This means that the stability condition for the process determining the number of these cars is satisfied, so our analysis will be based on the results obtained in section 1.6.

Let us move to the description of the models.

We consider an automobile road with two directions of traffic and m traffic lanes in each. The flow of cars X_i in the ith direction is a regenerative flow with intensity $\lambda_X^{(i)} (i = 1, 2)$. The road is intersected by a two-directional pedestrian crossing (Figure 1.1). We denote that pedestrians following from A to B have the first type, and from B to A have the second type. The flow of pedestrians of the ith type is a Poisson flow with intensity λ_i $(i = 1, 2)$. Pedestrians cross the road independently of each other with a random (but constant during the entire time of being at the crosswalk) speed.

First, we assume that there is no traffic light at the crossing and pedestrians have an absolute priority over cars. In this case the number of pedestrians at the crosswalk is the number of customers in the infinite-channel service system of $M|G|\infty$ type.

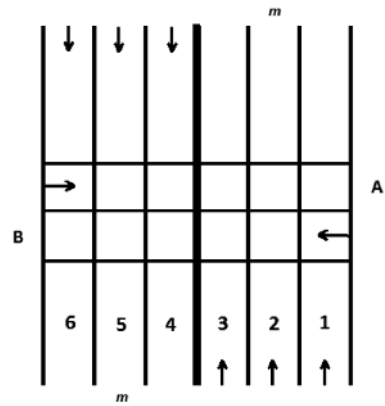

Figure 1.1. *A road intersected by a pedestrian crossing*

Let us assume that $2b$ is an average time of crossing the road by a pedestrian. Then the probability P_0 that there is no pedestrian at the crosswalk in a stationary regime is defined by the expression

$$P_0 = e^{-2(\lambda_1+\lambda_2)b} \qquad [1.22]$$

Saaty (1961).

First assume that a car can cross a pedestrian crossing only if there are no pedestrians on it. Let us assume that the cars in the lanes $1, 2, \ldots, m$ are going in one direction and the cars in the lanes $m+1, \ldots, 2m$ – in another. We consider the process $Q_1(t)$ – the number of cars in the lanes $1, 2, \ldots, m$ at time t (the consideration of lanes $m+1, m+2, \ldots, 2m$ is analogous).

Denote $H_j(t)$ the mathematical expectation of the number of cars that pass through the crosswalk at the lane j during time t under the condition that there are always cars at this lane and the crosswalk is free. Also denote $H(t) = \sum_{j=1}^{m} H_j(t)$. In relation to the process $Q_1(t)$, we have a single-channel service system with an unreliable server. The operating time u_1 has an exponential distribution with the parameter $\lambda_1 + \lambda_2$, and the unavailable time u_2 is the period of the system $M|G|\infty$ being busy.

Since

$$P_0 = \frac{Eu_1}{Eu_1 + Eu_2},$$

then

$$a = Eu_1 + Eu_2 = (\lambda_1 + \lambda_2)^{-1} e^{2(\lambda_1+\lambda_2)b}.$$

It is not difficult to show that under the assumptions made, the results of section 1.6 are correct and the traffic rate ρ_1 is determined by the expression

$$\rho_1 = \frac{\lambda_X^{(1)} e^{2(\lambda_1+\lambda_2)b}}{(\lambda_1+\lambda_2)h(\lambda_1+\lambda_2)}, \qquad [1.23]$$

where $h(\lambda) = \lambda \int_0^\infty e^{-\lambda y} H(y)dy$.

The necessary and sufficient condition for the stability of the process Q_1 is the fulfillment of the inequality $\rho_1 < 1$, and the capacity $\bar{\lambda}_X^{(1)}$ is defined as

$$\bar{\lambda}_X^{(1)} = (\lambda_1+\lambda_2)h(\lambda_1+\lambda_2)e^{-2(\lambda_1+\lambda_2)b}.$$

If, for example $H(t) = m\nu t$, which corresponds to the assumption that each car crosses the pedestrian crossing during an exponentially distributed time with a parameter ν, then

$$\bar{\lambda}_X^{(1)} = m\nu e^{-2(\lambda_1+\lambda_2)b}.$$

When the real intensity $\lambda_X^{(1)}$ is less than, but close to $\bar{\lambda}_X^{(1)}$, large queues accumulate before the crosswalk.

Their asymptotic analysis, as well as some results concerning characteristics of the process Q_1 in a stationary regime, when $\rho^{(1)} < 1$ can be found in the papers (Afanasyeva and Rudenko 2012; Afanasyeva and Mihaylova 2015).

Now we will consider model 2, in which the rules for crossing the crosswalk by a car are weakened. We assume that the car can move along the jth lane ($j = \overline{1,m}$) if there are no pedestrians of the first type (going from A to B) on the lanes $1, 2, \ldots, j$, and there are no pedestrians of the second type on the lanes $j, j+1, \ldots, 2m$. Denote $P_0(j)$ as the probability of this event in a steady-state.

Since the number of pedestrians of the first type on the lanes $(1, 2, \ldots, j)$ is the number of customers in the system $M|G|\infty$ with the intensity λ_2 and with an average service time $(2m-j+1)\frac{b}{m}$, then

$$P_0(j) = e^{-j\lambda_1 \frac{b}{m} - (2m-j+1)\lambda_2 \frac{b}{m}}. \qquad [1.24]$$

So we have a queueing system with m unreliable servers. All servers break when a pedestrian of the first (second) type appears on lane 1 (the $2m$th).This means that

the available time $\tau_j^{(1)}$ of the jth server is exponentially distributed with the parameter $\lambda_1 + \lambda_2$. Let $\tau_j^{(2)}$ be a block time of the jth server and $a_j = E\tau_j^{(1)} + E\tau_j^{(2)}$. Then $P_0(j) = \frac{E\tau_j^{(1)}}{a_j}$, so

$$a_j = (\lambda_1 + \lambda_2)^{-1} e^{j\lambda_1 \frac{b}{m} + (2m-j+1)\lambda_2 \frac{b}{m}}.$$

Assuming that $h_j(\lambda) = \lambda \int_0^\infty e^{-\lambda t} H_j(t)dt$ and using the results of section 1.6, we can find the traffic rate ρ_2 for model 2.

$$\rho_2 = \lambda_X^{(1)} [(\lambda_1 + \lambda_2) \sum_{j=1}^{m} P_0(j) h_j(\lambda_1 + \lambda_2)]^{-1}. \qquad [1.25]$$

If $h_j(\lambda) = \frac{1}{m} h(\lambda), j = \overline{1,m}$, then [1.25] can be written as

$$\rho_2(m) = \begin{cases} \frac{m\lambda_X^{(1)}}{(\lambda_1+\lambda_2)h(\lambda_1+\lambda_2)} \frac{e^{\lambda_2 \frac{b}{m}} - e^{\lambda_1 \frac{b}{m}}}{e^{-(\lambda_1+\lambda_2)b} - e^{-2\lambda_2 b}}, \text{ when } \lambda_1 \neq \lambda_2 \\ \frac{\lambda_X^{(1)} e^{2\lambda b + \frac{\lambda b}{m}}}{2\lambda h(2\lambda)}, \text{ when } \lambda_1 = \lambda_2. \end{cases}$$

When $m = 1$, we get

$$\rho_2(1) = \frac{\lambda_X^{(1)} e^{2\lambda_2 b + \lambda_1 b}}{(\lambda_1 + \lambda_2)h(\lambda_1 + \lambda_2)} < \frac{\lambda_X^{(1)} e^{2(\lambda_2 + \lambda_1)b}}{(\lambda_1 + \lambda_2)h(\lambda_1 + \lambda_2)} = \rho_1.$$

It is easy to show that for all $m \geq 1$, the inequality $\rho_2(m) < \rho_1$ holds. Weakening the rules of crossing the crosswalk increases the capacity of the route. To estimate this effect, we consider the ratio

$$\frac{\rho_2(m)}{\rho_1} = \begin{cases} \frac{m(e^{\lambda_2 \frac{b}{m}} - e^{\lambda_1 \frac{b}{m}})}{e^{(\lambda_1+\lambda_2)b} - e^{2\lambda_1 b}}, \text{ when } \lambda_1 \neq \lambda_2 \\ e^{-2\lambda b + \frac{\lambda b}{m}}, \text{ when } \lambda_1 = \lambda_2 = \lambda \end{cases}.$$

Putting $x = e^{\lambda_1 b} \geq 1$, $\lambda_2 = \alpha\lambda_1$, we have

$$\phi(x) = \frac{\rho_2(m)}{\rho_1} = \begin{cases} \frac{m(x^{\frac{\alpha}{m}} - x^{\frac{1}{m}})}{x^{1+\alpha} - x^2}, \; when\; \alpha \neq 1 \\ x^{-2+\frac{1}{m}}, \; when\; \alpha = 1 \end{cases}.$$

After drawing the graphs for $\phi(x)$ ($\phi(1) = 1$) for $\alpha = 0.5; 1.5; 2$ (see Figure 1.2), we can see that the effect of the weakened rule (model 2) in comparison with the standard rule (model 1) is stronger, as the number of the lanes increases and the intensity of the number of pedestrians increases.

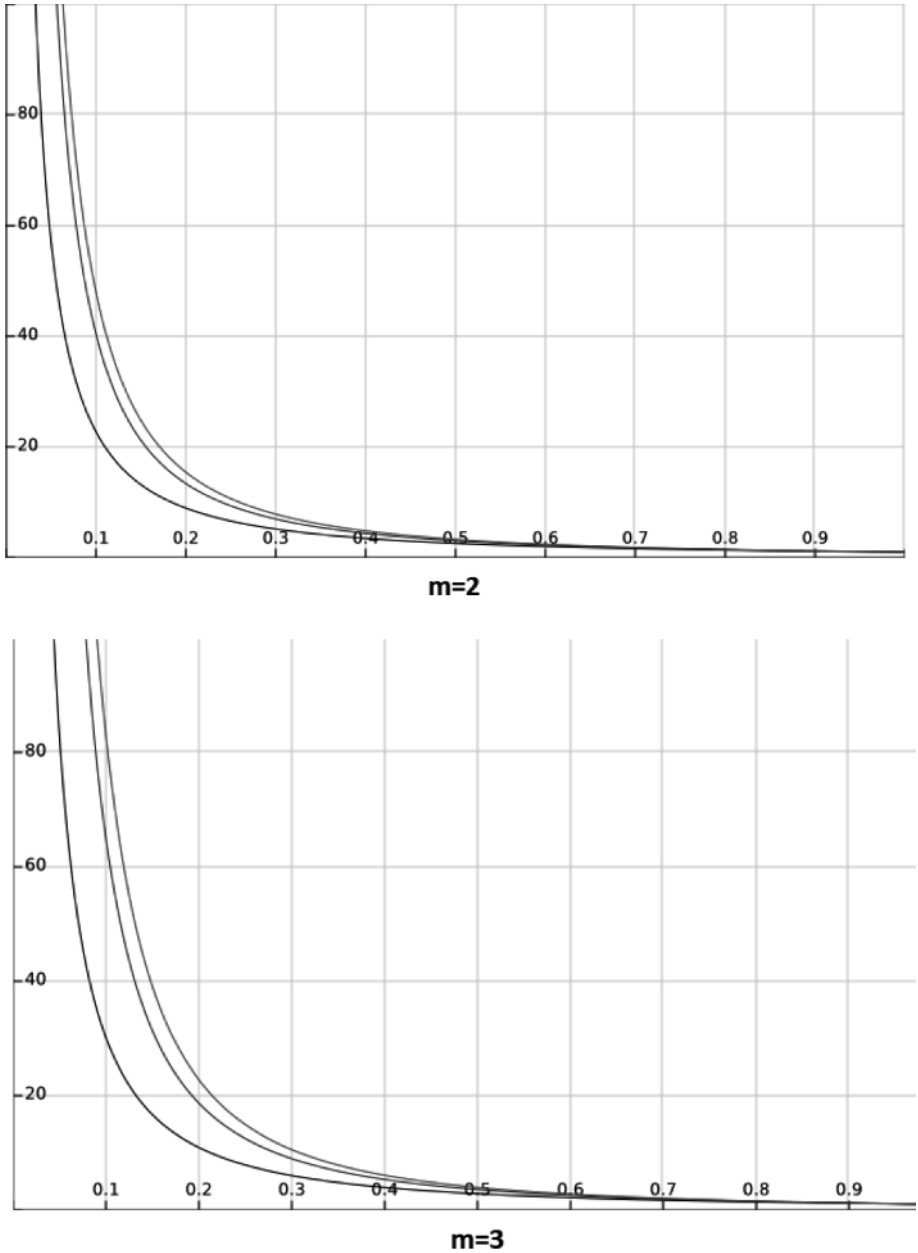

Figure 1.2. *Plots for* $\alpha = 0.5, 1.5, 2$

Currently there is no algorithm that estimates the number of cars before the crosswalk in model 2, however we can obtain asymptotic expressions for the average number of expected cars when $\rho_2 \uparrow 1$. It turns out that $EQ \sim \frac{c}{1-\rho_2}$, where c is a constant.

If the length of the queue of cars is unacceptably high, it is necessary to make organizational decisions. One of these decisions is to install a traffic light. Then, in relation to the cars, we again get a one-channel service system with an unreliable server, but now the server will not work if the red light is on (for cars) and will work if the green light is on. This model has been studied in papers (Afanasyeva and Bulinskaya 2013, 2010), in which the algorithms for estimating the queue length were proposed and the number of the asymptotic results were received. It can happen that, with the available traffic intensities of the cars and the pedestrians, the installation of a traffic light, even at the optimum interrelationship between switching intervals, does not provide an acceptable level of queues of pedestrians and cars. This may be used as the basis for the construction of an underground (or overground) pedestrian crossing, its elimination or transfer to another location.

1.11. Conclusion

In this chapter, we considered the stability problem for multiserver queues with a regenerative input flow. Let us note that stability analysis is an essential and challenging stage of the investigation of stochastic models. However, stability conditions may be of independent interest. In particular, the stability criterion of the multiserver model can be used for the capacity planning of a model and estimation of the upper bound of potential energy saving. The main contribution of our research is an extension of the stability criterion to the model with a regenerative input flow. The method we use has the following steps. First, we define an auxiliary process Y that is the number of customers that are served up to time t at the auxiliary system S_0 in which there are always customers for service. Second, assuming that this process is a regenerative flow not depending on the input flow X under some additional conditions, we construct the common points of regeneration of Y and X. We call this step synchronization of the processes. This approach allows us to use results from the renewal theory for the stability analysis of the systems satisfying additional conditions. These conditions may seem too restrictive to be useful for the analysis of the real models. Therefore, we apply our approach to the stability analysis of two classical systems with service interruptions (sections 1.6 and 1.7) of the queueing system with preemptive priority discipline (section 1.8) and of the system with simultaneous service of a customer by a random number of servers (section 1.9). It follows from our results that the structure of the input flow does not affect the stability condition. We only need to know the intensity of this flow to estimate the traffic rate. But for the preemptive repeat different service discipline, the distribution of the service time plays an essential role, since the traffic rate is expressed with the

help of the renewal function corresponding to this distribution. We obtain the upper bound for the traffic rate providing the sufficient stability condition. Note that this condition is the same as the condition obtained in Morozov *et al.* (2011) for a more simple model. Finally, in section 1.9 we give applications of our results for the estimation of the capacity of the automobile road, intersected by a crosswalk.

1.12. Acknowledgment

This work is partially supported by RFBR grant 17-01-00468.

1.13. References

Afanaseva, L. (2019). Asymptotic analysis of queueing models based on synchronization method. *Methodology and Computing in Applied Probability*, 1–22, https://doi.org/10.1007/s11009-019-09694-9.

Afanasyeva, L.G., Bashtova E.E. (2014). Coupling method for asymptotic analysis of queues with regenerative input and unreliable server. *Queueing Systems*, 76(2), 125–147.

Afanasyeva, L., Bashtova, E., Bulinskaya, E. (2012). Limit theorems for semi-Markov queues and their applications. *Communications in Statistics Part B: Simulation and Computation*, 41(6), 688–709.

Afanasyeva, L.G., Bulinskaya, E.V. (2009). Some problems for the flow of interacting particles. *Modern Problems of Mathematics and Mechanics*, 2, 55–68.

Afanasyeva, L.G., Bulinskaya, E.V. (2010). Mathematical models of transport systems based on queueing systems methods. *Proceedings of Moscow Institute of Physics and Technology*, 2(4), 6–21.

Afanasyeva, L.G., Bulinskaya, E.V. (2011). Stochastic models of transport flows. *Commun. Stat. Theory Methods*, 40(16), 2830–2846.

Afanasyeva, L.G., Bulinskaya, E.V. (2013). Asymptotic analysis of traffic lights performance under heavy-traffic assumption. *Methodology and Computing in Applied Probability*, 15(4), 935–950.

Afanasyeva, L.G., Mihaylova, I.V. (2015). Two models of the highway intersected by a crosswalk. *Survey of Applied and Industrial Mathematics*, 22(5), 520–532.

Afanasyeva, L.G., Rudenko, I.V. (2012). $GI|G|\infty$ queueing systems and their applications to the analysis of traffic models. *Theory of Probability. Applications*, 57(3), 427–452.

Afanasyeva, L., Tkachenko, A. (2014). Multichannel queueing systems with regenerative input flow. *Theory of Probability and Its Applications*, 58(2), 174–192.

Afanasyeva, L., Tkachenko, A. (2016). Stability analysis of multi-server discrete-time queueing systems with interruptions and regenerative input flow. In *New Trends in Stochastic Modeling and Data Analysis*, Manca, R., McClean, S., Skiadas, C.H. (eds). ISAST: International Society for the Advancement of Science and Technology, Athens, 13–26.

Afanasyeva, L., Tkachenko, A. (2018). Stability of discrete multi-server queueing systems with heterogeneous servers, interruptions and regenerative input flow. *Reliability: Theory and Applications*, 13(1), 63–75.

Asmussen, S. (2003). *Applied Probability and Queues*. Springer-Verlag, New York.

Avi-Itzhak, B., Naor, P. (1963). Some queueing problems with the service station subject to breakdown. *Operations Research*, 11(3), 303–320.

Baycal-Gursoy, M., Xiao, W. (2004). Stochastic decomposition in $M|M|\infty$ queues with Markov-modulated service rates. *Queueing Systems*, 48, 75–88.

Baycal-Gursoy, M., Xiao, W., Ozbay, K. (2009). Modeling traffic flow interrupted by incidents. *Eur. J. Oper. Res.*, 195, 127–138.

Belorusov, T. (2012). Ergodicity of a multichannel queueing system with balking. *Theory of Probability and Its Applications*, 56(1), 120–126.

Blank, M. (2003). Ergodic properties of a simple deterministic traffic flow model. *J. Stat. Phys.*, 111, 903–930.

Borovkov, A.A. (1976). *Stochastic Processes in Queueing Theory*. Springer-Verlag, New York.

Caceres, F.C., Ferrari, P.A., Pechersky, E. (2007). A slow to start traffic model related to a $M|M|1$ queue. *J. Stat. Mech.* arXiv:cond-mat/0703709 v2 [cond-mat.stat-mech].

Chen, H. (1995). Fluid approximation and stability of multiclass queueing networks: Work-conserving disciplines. *Annals of Applied Probability*, 5, 637–665.

Chen, H., Yao, D. (2001). *Fundamentals of Queueing Networks*. Springer, New York.

Chowdhury, D. (1999). Vehicular traffic: A system of interacting particles driven far from equilibrium. arXiv:arXiv:cond-mat/9910173 v1 [cond-mat.stat-mech].

Dai, J. (1995). On positive Harris recurrence of multiclass queueing networks: A unified approach via fluid limit models. *Annals of Applied Probability,* 5, 49–77.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, 2nd ed. John Wiley & Sons, New York.

Fiems, D., Bruneel, H. (2013). Discrete-time queueing systems with Markovian preemptive vacations. *Mathematical and Computer Modeling*, 57(3-4), 782–792.

Foss, S., Konstantopoulos, T. (2004). An overview on some stochastic stability methods. *Journal of the Operations Research Society of Japan*, 47(4), 275–303.

Fuks, H., Boccara, N. (2001). Convergence to equilibrium in a class of interacting particle system evolving in discrete time. *Phys. Rev. E.*, 64, 016117.

Gaver Jr., D. (1962). A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24, 73–90.

Georgiadis, L., Szpankowski, W. (1992). Stability of token passing rings. *Queueing Systems*, 11, 7–33.

Gideon, R., Pyke, R. (1999). Markov renewal modeling of Poisson traffic at intersections having separate turn lanes. In *Semi-Markov Models and Applications*, Janssen, J., Limneos, N. (eds). Springer, New York, NY, 285–310.

Gillent, F., Latouche, G. (1983). Semi-explicit solution for $M|PH|1$ – like queueing systems. *European Journal of Operational Research*, 13(2), 151–160.

Grandell, J. (1976). *Double Stochastic Poisson Process*, Lecture Notes in Mathematics, 529, Springer, Berlin.

Greenshields, B.D. (1935). A study of highway capacity. *Proc. Highway Res.*, 14, 448–477.

Grinbeerg, H. (1959). An analysis of traffic flows. *Oper. Res.*, 7, 79–85.

Helbing, D. (2001). Traffic and related self-driven many-particle systems. *Rev. Mod. Phys.*, 73, 1067–1141.

Inose, H., Hamada, T. (1975). *Road Traffic Control*. University of Tokyo Press, Tokyo.

Kiefer, J., Wolfowitz, J. (1955). On the theory of queues with many servers. *Trans. Amer. Math. Soc.*, 78, 1–18.

Krishnamoorthy, A., Pramod, P., Chakravarthy, S. (2012). Queues with interruptions: A survey. TOP, 1-31 doi:10.1007/s11750-012-0256-6.

Loynes, R.M. (1962). The stability of a queue with non-independent inter-arrival and service times. *Proc. Cambr. Phil. Soc.*, 58(3), 497–520.

Maerivoet, S., de Moor, B. (2005). Cellular automata models of road traffic. *Phys. Rep.*, 419, 1–64.

Malyshev, V.A., Menshikov, M.V. (1982). Ergodicity continuity and analyticity of countable Markov chains. *Trans Moscow Math*, 1, 1–48.

Meyn, S.P., Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, New York.

Morozov, E. (2004). Weak regeneration in modeling of queueing processes. *Queueing Systems*, 46, 295–315.

Morozov, E. (2007). A multiserver retrial queue: Regenerative stability analysis. *Queueing Systems,* 56(3-4), 157–168.

Morozov, E., Dimitriou, I. (2017). Stability analysis of a multiclass retrial system with coupled orbit queues. In *Computer Performance Engineering. EPEW 2017*, Reinecke, P., Di Marco, A. (eds). Lecture Notes in Computer Science. Springer, Cham, 85–98.

Morozov, E., Fiems, D., Bruneel, H. (2011). Stability analysis of multiserver discrete-time queueing systems with renewal-type server interruptions. *Performance Evaluation*, 68(12), 1261–1275.

Morozov, E., Rumyantsev, A. (2016). Stability analysis of a $MAP|M|s$ cluster model by matrix-analytic method. *European Workshop on Computer Performance Engineering*, 63–76.

Neuts, M. (1989). *Structured Stochastic Matrices of M/G/1 Type and Their Applications*. Marcel Dekker, New York.

Pechinkin, A., Socolov, I., Chaplygin, V. (2009). Multichannel queueing system with refusals of servers groups. *Informatics and Its Applications*, 3(3), 4–15.

Rumyantsev, A., Morozov, E. (2017). Stability criterion of a multi-server model with simultaneous service. *Annals of Operations Research*, 252(1), 29–39.

Saaty, T.L. (1961). *Elements of Queueing Theory with Applications*. McGraw-Hill, Inc, New York, NY.

Sadowsky, J.S. (1995). The probability of large queue lengths and waiting times in a heterogeneous multiserver queue: positive recurrence and logarithmic limits. *Adv. Appl. Prob.*, 27, 567–583.

Schadschneider A. (2000). Statistical physics of traffic flow. arXiv:arXiv:cond-mat/0007418 v1 [cond-mat.stat-mech].

Smith, W. (1955). Regenerative stochastic processes. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 232(1188), 6–31.

Szpankowski, W. (1994). Stability conditions for some distributed systems. Buffered random access systems. *Adv. Appl. Prob.*, 26, 498–515.

Thiruvengadam, K. (1963). Queueing with breakdowns. *Operations Research*, 11(1), 62–71.

Thorisson, H. (2000). *Coupling, Stationary and Regeneration*. Springer, New York.

White, H., Christie, L.S. (1958). Queueing with preemptive priorities or with breakdown. *Operations Research*, 6(1), 79–95.