

THE STABILITY OF LOAD BALANCED NETWORKS WITH GENERAL INPUT AND OUTPUT PROCESSES

VYACHESLAV M. ABRAMOV

ABSTRACT. The paper establishes necessary and sufficient conditions for stability of load-balanced networks under quite general settings on arrival and departure processes. The approach of this paper is new and simpler than those known earlier. It matches networks under more general assumptions than before. For generalized Jackson networks, which are a particular case of those, the proofs become very elementary. It is shown in the paper, that the necessary and sufficient condition for stability of load-balanced networks is related to the solution of a linear programming problem precisely formulated in the paper.

1991 *Mathematics Subject Classification*. Primary: 60K25, 90B15. Secondary: 90C05.

Key words and phrases. Parallel queues, Join-the-shortest-queue models, Load-balanced network, Point Processes, Stability, Linear programming.

1. INTRODUCTION

1.1. **The goal of the paper.** The parallel queueing systems are well-presented in the literature. They have good properties (e.g. [17], [24], [31], [39], [50]) resulting in various applications in science and technology.

In the present paper, we study the stability of join-the-shortest-queue models including load-balancing networks. There are a number of works in the literature devoted to this problem (e.g. [15], [18], [19], [20], [30], [42], [45], [46], [47] and [48].) The goal of the present paper is to solve stability problem under possibly general framework (arrival processes are point processes rather than renewal processes, arrivals and departures can be dependent and so on), which are later described in the paper.

1.2. **Motivation.** The stability of stochastic processes and especially queueing systems and networks of queues is a distinguished area of research, and there is a large number of books and research papers devoted to the stability. The aim of this section is to motivate the method by showing the relevant chronology rather than give a review of this area. The detailed review can be found, for example, in the recent paper/book of Bramson [11], [12].

The first paper on stability of queueing systems assuming dependence of interarrival and service times was due to Loynes [32]. It led to a new vision of the stability for models with *dependent* interarrival and/or service times, and has been a source for new methods of the stability for complex queueing models. Among them there are renovative theory and recurrence equation methods (e.g. [7], [13] and many others.)

Further development of the Loynes approach towards queueing networks, because of their complicated nature, has been problematical, and the proof of the stability and ergodicity of queueing networks is typically based on other methods based on special theories of Markov and regeneration processes.

The first results on the stability of Jackson-type queueing networks have been obtained by Borovkov [8], [9]. Then the stability and ergodicity of networks have been studied in many papers. We mention the papers by Meyn and Down [34], Kaspi and Mandelbaum [26], [27], Sigman [43] and Baccelli and Foss [5]. All

of these papers but [5] are based on regeneration phenomena of the theory of Harris recurrent Markov chains.

The theory of Harris recurrent Markov chains has been exposed in the books of Orey [40] and Revuz [41]. The detailed study of stochastic stability of Markov chains and the theory of Harris recurrent Markov processes can be found in the book of Meyn and Tweedie [38], and in a number of research papers of these authors [35], [36] and [37]. In the book of Borovkov [10] the Harris recurrence of Markov chains is explained in the framework of more general recurrence sequences.

However, the proof of networks stability by the means of Harris recurrent Markov chains is restrictive. It mainly works in the cases where the sequences of interarrival and service times consist of independent and identically distributed random variables. In this case the phase space of the process can be expanded to Markov, and a network stability is proved in terms of the stability of the corresponding Markov process. As a rule the proof of the network stability in this case requires special additional conditions. For example, in most of the papers the infinite support of interarrival time distributions as well as some additional method based technical conditions are required (e.g. see Dai [14]).

Baccelli and Foss [5] proved the stability of Jackson-type networks with dependent interarrival and service times. Their proof is based on development of renovation theory. However, the mentioned paper is about 70 pages long, it is conceptually difficult and based on the results from different areas (stochastic Petri nets for example).

In the present paper, new conditions for the stability of JS-queue models including load balanced networks are established. The class of load-balanced networks is wider than the class of Jackson-type networks, so the stability results of the present paper are more general than those for Jackson-type networks. On the other hand, our method is a Loynes-based method, and our stability results are established for quite general networks with sequences of *dependent* interarrival times. Service times can be dependent, and the sequences of interarrival and service times can be dependent of each other as well.

In addition, our results are obtained under weaker conditions than the known results. Stationarity of the arrival point processes is not assumed. The only weaker condition of the strong law of large number for the sequences of inter-arrival times is used.

Our challenge is as follows. Using sample-path techniques we first establish the equivalence: *the stability of usual queueing systems follows from the stability of queueing systems with an autonomous service mechanism*. Then we establish stability of queueing systems and networks with an autonomous service mechanism, which is a much easier problem and can be solved under quite general assumptions with the aid of the Loynes-based method.

A sample-path proof of the correspondence between the stability of usual queueing systems and queueing systems with an autonomous service mechanism is clear and elementary. It results in a significant contribution to the overall elementary proof of the stability of different queueing networks. A sample-path approach for the stability of queueing systems and networks is well-known in the literature (e.g. [16]). The novelty of the approach of the present paper is to use a sample-path analysis in combination with other methods. It is based on a new idea of a reduction of the original problem to another not traditional and simpler problem.

Queueing systems with an autonomous service mechanism have been introduced and initially studied by Borovkov [6], [7], and then have been an object of study in a large number of papers (e.g. [1], [2], [4], [21], [22], [23]). In traditional applications, queueing systems with autonomous service are associated with a shuttle bus picking up passengers from stations. Nowadays, however, there are many new examples arising out of computer technologies, where queueing systems with an autonomous service mechanism can be applied.

As was mentioned, the idea of this paper is to reduce one stability problem of a complex network to another stability problem of a network with simpler/concise properties. This idea is not new. For example, there are special criteria allowing to replace stability problem of an original *stochastic* network by stability problem of *deterministic* fluid model (see e.g. Dai [14]). Although

it looks natural to reduce the original problem related to stochastic network by the other similar problem related to deterministic fluid model, it requires additional (mild) conditions. The advantage of our sample-path analysis method is that no additional technical conditions is required. As well, it follows from our general method that the network to be stable in the terms of this paper need not be Harris recurrent.

It is worth noting, that the similar idea of comparing an original multiserver system with its analog with an autonomous service mechanism has been used in the theory of heavy traffic analysis by Iglehart and Whitt [25] in their paper of 1970. The approach of the present paper is simpler. We compare single server queueing systems, and our sample-path comparison is based on simple and intuitively understandable two-sided sample-path inequalities.

1.3. Organization of the paper. The rest of the paper is organized as follows. In Section 2 we describe main JS-queue models, which will be then developed and modified in the following sections. (The material of the paper is presented in the order of increasing complexity.) In the same section we give all necessary definitions related to the stability of queues and networks. In Section 3 we prove the correspondence between the stability of the original queueing system and that of the associated queueing system with an autonomous service mechanism. The proof is based on sample path analysis. In Section 4 we establish conditions for stability of JS-queue models of queueing systems, and then in Section 5 we establish conditions for the stability of load-balanced networks. In Section 6 we conclude the paper, where the stability of more general networks, than those studied here, with batch arrival and service times are discussed.

2. DESCRIPTION OF MAIN MODELS AND DEFINITIONS OF THE STABILITY

2.1. Main models. In this section we describe main JS-queue models with an autonomous service mechanism. These models and some of the assumptions related to the arrival and departure processes will be then modified in the following sections.

- There are m identical servers, each of which having its own queue.

- All of the processes that describe queueing models are assumed to be right-continuous and to have left limits.

- The arrival process is governed by two point processes $A(t)$ and $A'(t)$. The process $A(t)$ is defined by a sequence $\{\tau_n\}_{n \geq 1}$ of positive random variables, and the corresponding sequence of points is as follows: $t_1 = \tau_1$, and $t_{n+1} = t_n + \tau_{n+1}$, $n \geq 1$. Then, $A(t) = \sum_{i=1}^{\infty} \mathbf{1}_{\{t_i \leq t\}}$. The process $A'(t)$ is defined analogously. We have the sequence of positive random variables $\{\tau'_n\}_{n \geq 1}$, and the sequence of points $t'_1 = \tau'_1$, and $t'_{n+1} = t'_n + \tau'_{n+1}$, $n \geq 1$. Then, $A'(t) = \sum_{i=1}^{\infty} \mathbf{1}_{\{t'_i \leq t\}}$. We assume

$$\mathbf{P} \left\{ \lim_{t \rightarrow \infty} \frac{A(t)}{t} = \lambda \right\} = 1, \quad (2.1)$$

and

$$\mathbf{P} \left\{ \lim_{t \rightarrow \infty} \frac{A'(t)}{t} = \lambda' \right\} = 1. \quad (2.2)$$

The process $A(t)$ forms a *dedicated* traffic, while the process $A'(t)$ forms an *opportunistic* traffic.

- A customer arriving at moment t_n , $n \geq 1$, is assigned to the j th queue, $j = 1, 2, \dots, m$, with the probability p_j ($\sum_{j=1}^m p_j = 1$), residing there to wait for the service.

- A customer, arriving at moment t'_n , $n \geq 1$, is assigned to the queue with the shortest queue-length, where equal probability tie-breaks are assumed if there are several shortest queue-lengths.

- The departure process from the j th server is governed by the renewal process $D^{(j)}(t)$. These renewal processes are mutually independent and have the same distribution for all j . Specifically, the n th service time of the j th server is denoted $\chi_n^{(j)}$, and the corresponding sequence of points is denoted $\{x_n^{(j)}\}$ where $x_1^{(j)} = \chi_1^{(j)}$ and $x_{n+1}^{(j)} = x_n^{(j)} + \chi_{n+1}^{(j)}$, $n \geq 1$. We assume that $E\chi_n^{(j)} = \frac{1}{\mu}$. This leads to the relation

$$\mathbf{P} \left\{ \lim_{t \rightarrow \infty} \frac{D^{(j)}(t)}{t} = \mu \right\} = 1. \quad (2.3)$$

- For our convenience we assume that the processes $A(t)$, $A'(t)$ and $D^{(j)}(t)$, $j = 1, 2, \dots, m$, all are mutually independent processes. (Then this condition

together with other conditions (2.1), (2.2) and (2.3) will be relaxed.) It is also assumed that $D^{(j)}(t)$, $j = 1, 2, \dots, m$ are renewal processes. This assumption, in fact, can be relaxed. For example, one can assume that each sequence of service times $\{\chi_1^{(j)}, \chi_2^{(j)}, \dots\}$, $j = 1, 2, \dots, m$, forms a Markov chain. (The coupling arguments, which are then used for the sequences of independent and identically distributed service times can be easily adapted for Markov chains. The coupling arguments for Markov chains are well-known, e.g. [28].) However, in this case the point processes $A(t)$, $A'(t)$ and $D^{(j)}(t)$ all are assumed to be mutually independent. We will show later that the models, where $D^{(j)}(t)$, $j = 1, 2, \dots, m$ can be point processes that depend on $A(t)$ and $A'(t)$, are also appropriate. Possible ways of relaxing the initial assumptions on the processes $D^{(j)}(t)$ will be explained in Section 3.6.

- For technical convenience, the processes $A(t)$, $A'(t)$ and $D(t)$ all are assumed to be disjoint, i.e. $\mathbf{P}\{t_i = t'_j\} = \mathbf{P}\{t_i = x_k\} = \mathbf{P}\{t'_j = x_k\} = 0$, for all $i, j, k \geq 1$.

- The service mechanism of each server is assumed to be autonomous. This means the following. Let $Q^{(j)}(t)$ denote the number of customers in the j th queue at time t , $j = 1, 2, \dots, m$, and let $Q^{(j)}(0) = 0$. Let $A^{(j)}(t)$ and $A'^{(j)}(t)$ denote the thinning of the processes $A(t)$ and $A'(t)$ respectively, where $A^{(j)}(t)$ and $A'^{(j)}(t)$ are arrival processes to the j th queue. Then,

$$Q^{(j)}(t) = A^{(j)}(t) + A'^{(j)}(t) - \int_0^t \mathbf{1}_{\{Q^{(j)}(s-) > 0\}} dD^{(j)}(s), \quad (2.4)$$

where $Q^{(j)}(s-) = \lim_{u \uparrow s} Q^{(j)}(u)$. For the further convenience the above model is denoted φ_m , where the subscript m denotes the number of parallel queues.

We will also consider the particular case of the model φ_m , where

$$p_1 = p_2 = \dots = p_m = \frac{1}{m}.$$

In this case the families $\{A^{(j)}(t)\}_{j \leq m}$ and $\{A'^{(j)}(t)\}_{j \leq m}$ are each constituted of identically distributed processes. The above symmetric model with m parallel queues is denoted Σ_m .

2.2. Definitions of the stability. Above equation (2.4) is given for all $t \geq 0$. For our purpose we extend this equation, assuming that all the processes start at a . Then, instead of (2.4) for all $t \geq a$ we have the following equation:

$$Q^{(j)}(t-a) = A^{(j)}(t-a) + A'^{(j)}(t-a) - \int_0^{t-a} \mathbf{1}_{\{Q^{(j)}(s-) > 0\}} dD^{(j)}(s). \quad (2.5)$$

Definition 2.1. The system \wp_m is said to be stable if there exists a bounded set \mathcal{S} such that

$$\limsup_{a \rightarrow -\infty} \mathbf{P} \left\{ Q^{(j)}(t-a) \in \mathcal{S} \right\} > 0$$

for all $j = 1, 2, \dots, m$ and any t .

This definition remains in force for all the JS-queue models included load-balanced networks considered in the paper.

Comment to the definition of stability. For these general processes the above definition of the stability is rougher than other known definitions of the stability, for example that in the theory of Harris recurrent Markov processes. The approach of the present paper, however, enables us to obtain a reduction from this definition to the definition in the theory of Harris recurrent Markov processes, when the processes $A(t)$, $A'(t)$ and $D^{(j)}(t)$, $j = 1, 2, \dots$ all are renewal (see Sections 4.3 and 5.2).

3. SAMPLE-PATH COMPARISON OF QUEUEING SYSTEMS

In this section we compare three different queueing systems given on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and therefore defined by the same governing sequences of random variables, but different specific rules of departure. For simplicity we assume that all of these three systems start at zero with an empty queue.

These systems are defined by an arrival point process $A(t)$ and a departure renewal process $D(t)$. These processes are defined by the corresponding governing sequences $\{\tau_n\}$ and $\{\chi_n\}$, where χ_1, χ_2, \dots are independent identically distributed random variables. Let $t_k = \tau_1 + \tau_2 + \dots + \tau_k$ and $x_k = \chi_1 + \chi_2 + \dots + \chi_k$.

Then, the point process $A(t)$ and the renewal process $D(t)$ are

$$A(t) = \sum_{i=1}^{\infty} \mathbf{1}_{\{t_i \leq t\}}, \quad D(t) = \sum_{i=1}^{\infty} \mathbf{1}_{\{x_i \leq t\}}.$$

3.1. Definition of the first queueing system. The first queueing system is a queueing system with an autonomous service mechanism, which is denoted by \mathcal{Q}_1 . The queue-length process for this system is defined as

$$Q_1(t) = A(t) - \int_0^t \mathbf{1}_{\{Q_1(s-) > 0\}} dD(s).$$

3.2. Definition of the second queueing system. The second queueing system is a *usual* queueing system denoted by \mathcal{Q}_2 . The queue-length process of \mathcal{Q}_2 is correspondingly denoted by $Q_2(t)$ and is defined by a well-known recurrence equation. Specifically, $Q_2(t)$ is defined by *interrupted governing sequences* (cf. [6]) as follows. Denote

$$\eta_1 = \sup\{k : t_1 + x_k < t_{k+1}\}.$$

Then for $0 \leq t < t_1$ we have $Q_2(t) = 0$, and for $t_1 \leq t < t_1 + x_{\eta_1}$ the queue-length $Q_2(t)$ is defined as difference between the number of arrivals and service completions during the interval $[t_1, t]$ including the arrival at the instant t_1 , i.e. $Q_2(t) = [A(t) - A(t_1-)] - [D(t) - D(t_1)]$. Then, for $t_1 + x_{\eta_1} \leq t < t_{\eta_1+1}$ we have $Q_2(t) = 0$. Next, let

$$\eta'_1 = \sup\{k \geq \eta_1 : t_1 + x_k < t_{\eta_1+1}\}.$$

Similarly to the above denote

$$\eta_2 = \sup\{k > \eta'_1 : t_{\eta_1+1} + x_k - x_{\eta'_1} < t_{k+1}\}.$$

Then for $t_{\eta_1+1} \leq t < t_{\eta_2+1}$ the queue-length $Q_2(t)$ is defined as difference between the number of arrivals and service completions during the interval $[t_{\eta_1+1}, t]$ including the arrival at the instant t_{η_1+1} , i.e. $Q_2(t) = [A(t) - A(t_{\eta_1+1}-)] - [D(t) - D(t_{\eta_1+1})]$.

The stopping times η_3, η_4, \dots are defined similarly.

3.3. Definition of the third queueing system. The third queueing system is a special queueing system with delayed departures is denoted \mathcal{Q}_3 . The queue-length process of this system is defined as follows. The arrival process $A(t)$ is the same as in the aforementioned systems \mathcal{Q}_1 and \mathcal{Q}_2 , but departures of the customers are delayed as follows. Considering the queueing system \mathcal{Q}_1 let us assume that all first potential departures after arrivals of customers to an empty system are deleted. Specifically, let $\omega_A(t)$ denote the event occurring under the following conditions:

(a) the last arrival before time t is to an empty system: let t_A denote this moment of arrival;

(b) during the time interval $[t_A, t)$ there is no point of the departure process $D(t)$. Namely,

$$\eta = \sup\{k : x_k < t_A\} \text{ and } x_{\eta+1} \geq t.$$

Then, the queue-length process $Q_3(t)$ is defined as follows:

$$Q_3(t) = A(t) - \int_0^t (1 - \mathbf{1}_{\omega_A(s)}) \mathbf{1}_{\{Q_3(s-) > 0\}} dD(s), \quad (3.1)$$

where $\mathbf{1}_{\omega_A(s)}$ is the indicator of the event $\omega_A(s)$.

For a more clarity, a fragment of departure moments of the system \mathcal{Q}_3 compared to those moments of the system \mathcal{Q}_1 with autonomous service is shown in Figure 1. The upper scale of this figure contains the moments of departures in the queueing system with an autonomous service mechanism. These departure moments are connected by a segment of line with the corresponding departure moments in the queueing system with delayed departures. The scale placed in the middle of Figure 1 indicates the points $x_1, x_2, \dots, x_n, x_{n+1}$ related to these departures. The ‘delete’ sign in the upper and lower scales shows an absence of real departure the systems when they are empty.

3.4. Sample path comparison of \mathcal{Q}_1 and \mathcal{Q}_3 at departure moments.

The following Proposition 3.1 is based on the definition of the queueing systems \mathcal{Q}_1 and \mathcal{Q}_3 .

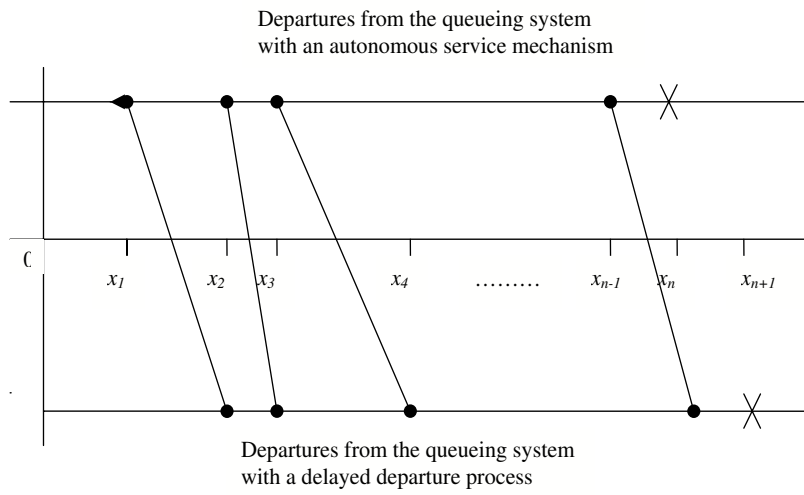


FIGURE 1. Correspondence between departure moments in \mathcal{Q}_1 (upper) and \mathcal{Q}_3 (lower).

Proposition 3.1.

$$Q_3(t, \omega) - Q_1(t, \omega) \leq 1. \quad (3.2)$$

Proof. For the purpose of the proof we will follow up the sample paths of the both processes of the queueing systems \mathcal{Q}_1 and \mathcal{Q}_3 .

Since all of the queue-length processes are defined on the same probability space, then in the further consideration they are provided by the additional argument $\omega \in \Omega$ in the places where it is required.

Apparently, that $Q_1(t, \omega) = Q_3(t, \omega) = 0$ for all $t \in [0, t_1)$ (recall that according to convention $A(0) = D(0) = 0$). For the system \mathcal{Q}_1 , let

$$l = \inf \{i : x_i > t_1\}.$$

Then, for any t from the interval $[t_1, x_l)$ we have $Q_1(t, \omega) = Q_3(t, \omega) = A(t, \omega)$, while in the point x_l itself we have $Q_3(x_l, \omega) - Q_1(x_l, \omega) = 1$, see Figure 2 (a).

Let v be a new number greater than l satisfying the property

$$v = \inf \{k > l : Q_1(x_k-) = 0\}.$$

(If such a number does not exist, then v equates to infinity. In this case, obviously, $Q_3(t, \omega) - Q_1(t, \omega) = 1$, $t \geq x_l$.) Then for all t of the interval $[x_l, x_v)$ we have $Q_3(t, \omega) - Q_1(t, \omega) = 1$, while in the point x_v itself we arrive at $Q_1(x_v, \omega) = Q_3(x_v, \omega) = 0$, see Figure 2 (b).

Thus, we arrived at zeroth queue-lengths again. The further paths of the both processes after point $t = x_v$ behave similarly to those after the point $t = 0$, i.e. the difference $Q_3(t, \omega) - Q_1(t, \omega)$ can take only one of the two values 0 or 1. \square

Following Proposition 3.1, we provide another definition of the queue-length process $Q_3(t)$ at the departure moments x_i , $i \geq 1$ ($x_0 = 0$) compared to them of the queue-length process $Q_1(t)$ assuming that both of these processes are considered on the same probability space. The definition, which is based on (3.1), is

$$Q_3(x_i-) = \begin{cases} Q_1(x_i-) + 1, & \text{if } Q_1(x_{i-1}-) > 0, \\ Q_1(x_i-), & \text{if } Q_1(x_{i-1}-) = 0 \end{cases} \quad (3.3)$$

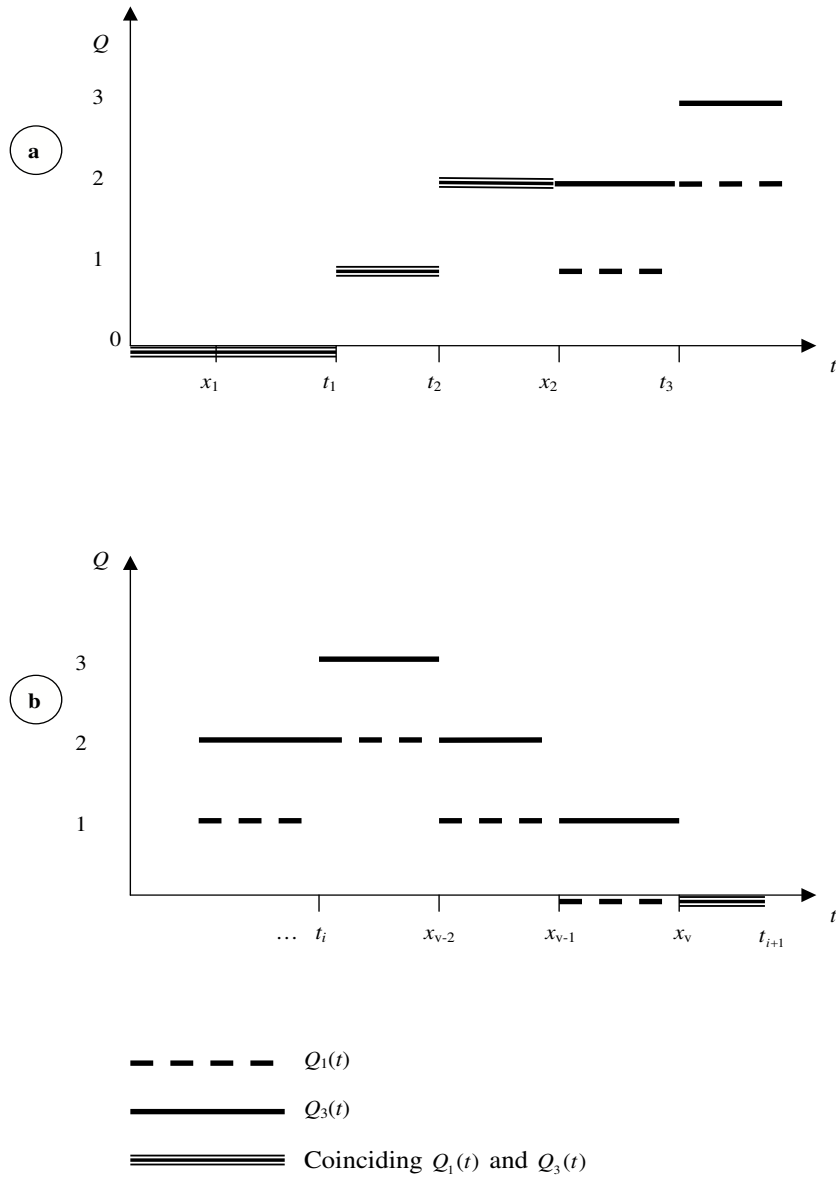


FIGURE 2. Typical sample paths of the queue-length processes $Q_1(t)$ and $Q_3(t)$:

(a) Sample paths of the queue-length processes after the origin;

(b) Sample paths of the queue-length processes before approach-

ing zero at point x_v .

Let us explain relation (3.3).

First, that the event $\{Q_1(x_{i-1}-) > 0\}$ in the first line of (3.3) occurs together with the event $\{Q_3(x_{i-1}-) > 0\}$. Moreover, $\{Q_1(x_{i-1}-) > 0\}$ means that either $\{Q_3(x_{i-1}-) = Q_1(x_{i-1}-)\}$ or $\{Q_3(x_{i-1}-) = Q_1(x_{i-1}-) + 1\}$. The first of these events occurs together with the event $\omega_A(x_{i-1})$. Then, at the moment x_{i-1} itself we have: $Q_3(x_{i-1}) = Q_3(x_{i-1}-)$ while $Q_1(x_{i-1}) = Q_1(x_{i-1}-) - 1$, i.e. $Q_3(x_{i-1}) = Q_1(x_{i-1}) + 1$, and consequently

$$Q_3(x_{i-}) = Q_1(x_{i-}) + 1. \quad (3.4)$$

If there occur the second event $\{Q_3(x_{i-1}-) = Q_1(x_{i-1}-) + 1\}$, then the event $\omega_A(x_{i-1})$ does not hold, and in the point x_{i-1} itself $Q_3(x_{i-1}) = Q_1(x_{i-1}) + 1$. Then we apparently arrive at (3.4). So, the first line of relation (3.3) is explained.

Let us explain the equation in the second line.

The complex event

$$\{Q_1(x_{i-1}-) = 0 \text{ and } Q_3(x_{i-1}-) > 1\}$$

is impossible according to Proposition 3.1.

The other complex event

$$\{Q_1(x_{i-1}-) = 0 \text{ and } Q_3(x_{i-1}-) = 1\}$$

occurs when the event $\omega_A(x_{i-1})$ does not hold. So, in the point x_{i-1} itself $Q_1(x_{i-1}) = Q_3(x_{i-1}) = 0$ (because only $Q_3(x_{i-1})$ is decremented in this point), and consequently both $Q_1(x_{i-})$ and $Q_3(x_{i-})$ are equal to the number of arrivals during the interval $[x_{i-1}, x_i]$.

In the case of the complex event

$$\{Q_1(x_{i-1}-) = 0 \text{ and } Q_3(x_{i-1}-) = 0\}$$

we again have $Q_1(x_{i-1}) = Q_3(x_{i-1}) = 0$ in the point x_{i-1} itself. The only difference from the case before is that $Q_3(x_{i-1})$ is not decremented in this point, because it is zero.

3.5. Coupling of the queue-length processes \mathcal{Q}_1 , \mathcal{Q}_2 and \mathcal{Q}_3 .

Proposition 3.2.

$$Q_1(t, \omega) \leq Q_2(t, \omega) \leq Q_3(t, \omega). \quad (3.5)$$

Proof. The proof of this proposition follows from a sample path comparison of queues in the queueing systems \mathcal{Q}_1 , \mathcal{Q}_2 and \mathcal{Q}_3 (see Figure 3). According to this comparison, the largest queue-length is in the queueing system \mathcal{Q}_3 (i.e. $Q_3(t, \omega) \geq Q_2(t, \omega)$ and $Q_3(t, \omega) \geq Q_1(t, \omega)$), because there is a delay for the service beginning of a customer arriving to an empty system. A customer arriving in an empty system \mathcal{Q}_2 is served without delay. In turn, the time elapsed from the moment of arrival of a customer in an empty \mathcal{Q}_1 system until its departure is shorter than the length of a genuine service time in the queueing system \mathcal{Q}_2 . \square

From Propositions 3.2 and 3.1 we arrive at the following conclusion: *if the queueing system \mathcal{Q}_1 is stable in the sense of Definition 2.1, then both of the queueing systems \mathcal{Q}_2 and \mathcal{Q}_3 are stable as well.* Following Proposition 3.2 this statement of stability can be extended for more complicated constructions including two, three and more nodes in the network and can be then applied to general Jackson-type networks.

Remark 3.3. The sample path comparison on the same probability space provided in this section is associated with the cases in the JS-queueing systems and networks considered in the following sections, where the processes $D^{(j)}(t)$ ($j = 1, 2, \dots, m$) all are independent and identically distributed renewal processes. If the processes $D^{(j)}(t)$ all are mutually independent *point* processes rather than renewal processes, then a sample path comparison is not correct in general. For example, if increments of a point process depend on time, then the realizations of increments, being shifted, become different from those initial. Similar situation can arise when the point processes $D^{(j)}(t)$ depend on the processes $A(t)$ or/and $A'(t)$. In some cases, however, one can assume that $D^{(j)}(t)$ are special point processes rather than renewal processes. The elementary examples and their extension is given in the section below.

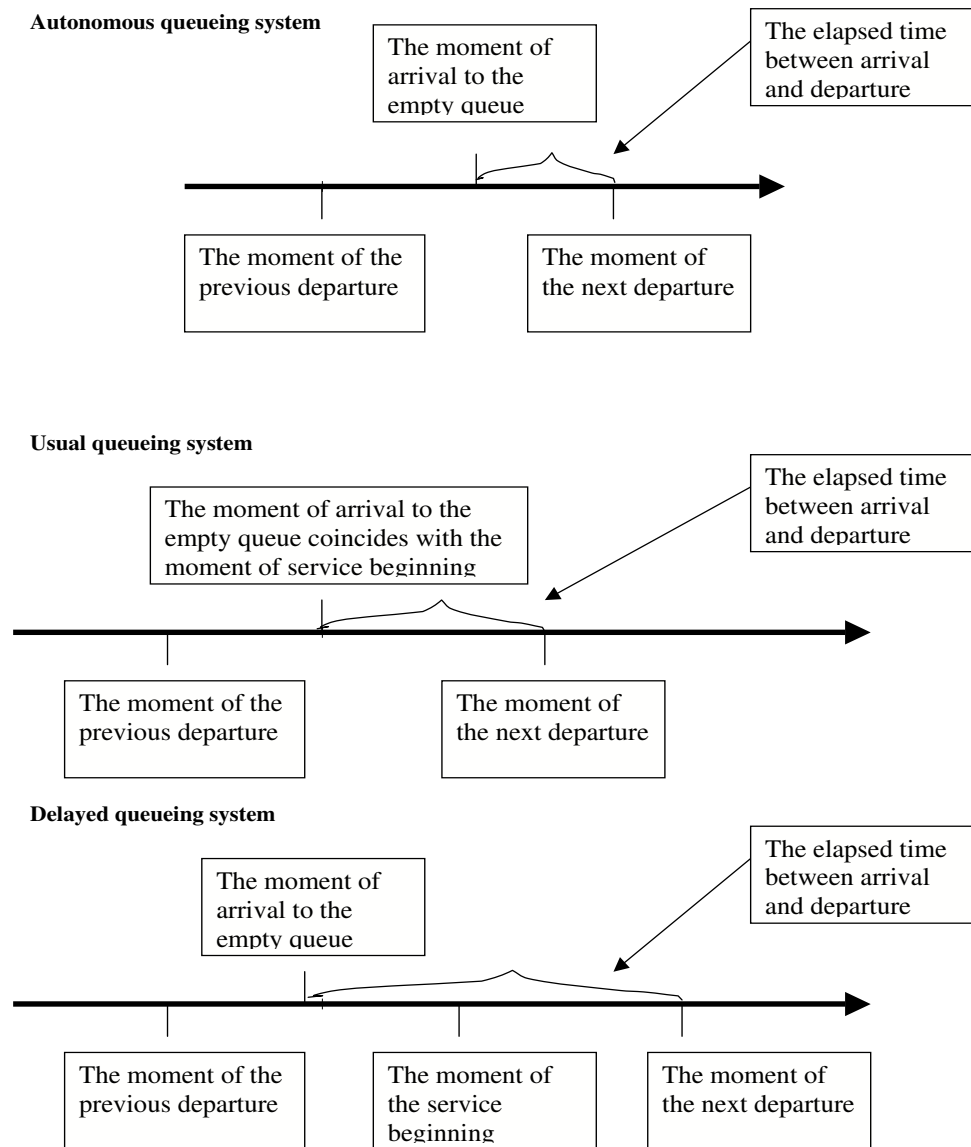


FIGURE 3. Behavior of the queueing systems \mathcal{Q}_1 , \mathcal{Q}_2 and \mathcal{Q}_3 when a queue is empty

3.6. Cases of queueing systems where $D(t)$ can be a point process. We start from two elementary examples. Example 1 is trivial and provided here for more clarity. Example 2 includes the case of dependent processes $D(t)$ and $A(t)$ and is crucial for further more extended constructions.

Example 1. The point process $D(t)$ is constructed as follows. The random variable χ_1 is uniformly distributed in the interval $[0, \frac{2}{\mu}]$, and $\chi_{i+1} = \frac{2}{\mu} - \chi_i$, $i \geq 1$.

Example 2. The similar construction can be provided for the point processes $D(t)$ depending on $A(t)$ as follows. Assume that χ_1, χ_3, \dots with odd order numbers are uniformly distributed in the interval $[0, \frac{2}{\mu}]$, and their realizations depend on the corresponding realizations of τ_1, τ_3, \dots . Then $\chi_{2i} = \frac{2}{\mu} - \chi_{2i-1}$, $i = 1, 2, \dots$

Proposition 3.4. *Propositions 3.2 and 3.1 remain true if $D(t)$ is a point process which is defined as follows. χ_1, χ_3, \dots with odd order numbers are uniformly distributed in the interval $[0, \frac{2}{\mu}]$, and generally depend on τ_1, τ_3, \dots ; $\chi_{2i} = \frac{2}{\mu} - \chi_{2i-1}$, $i = 1, 2, \dots$*

Proof. Introduce a new queueing system \mathcal{Q}_1^* with an autonomous service mechanism as follows. Assume that before the time instant t_1 the queue is empty, and customers arrive by batch of 2 at the instants

$$t_1 = \tau_1, t_3 = \tau_1 + \tau_2 + \tau_3, \dots, t_{2n-1} = \sum_{i=1}^{2n-1} \tau_i, \dots,$$

and, if the queue is not empty, then departures of batches containing those two customers occur at moments:

$$x_2 = \frac{2}{\mu}, x_4 = \frac{4}{\mu}, \dots, x_{2n} = \frac{2n}{\mu}, \dots$$

Using sample path arguments it is easily seen that there is the following relation between the queue length $Q_1^*(t, \omega)$ of this queueing system and the queue length $Q_1(t, \omega)$:

$$-1 \leq Q_1^*(t, \omega) - Q_1(t, \omega) \leq 2, \quad (3.6)$$

where the notation $Q_1(t, \omega)$ for the queue length is associated with the queueing system with the construction described in this proposition. Interarrival times are τ_1, τ_2, \dots , and the service times are as defined in this proposition.

Inequality (3.6) can be explained as follows. Let

$$\ell = \max \left\{ i : \sum_{k=1}^i \tau_k \leq \sum_{k=1}^i \chi_k \right\}.$$

This means that both $Q_1(t, \omega)$ and $Q_1^*(t, \omega)$ are positive for $\tau_1 \leq t < \chi_\ell$ and all $\omega \in \Omega$. The following two possible cases are as follows. (Below the argument ω is omitted.)

Case 1. ℓ is even. In this case $Q_1^*(\tau_{\ell+1}) = 2$, $Q_1(\tau_{\ell+1}) = 1$, and hence $Q_1^*(\tau_{\ell+1}) - Q_1(\tau_{\ell+1}) = 1$. After the time instant $\tau_{\ell+1}$ the difference $Q_1^*(t) - Q_1(t)$ can achieve 2 if a next event after time instant $\tau_{\ell+1}$ is only a departure from the queueing system \mathcal{Q}_1 , and nothing occurs in the queueing system \mathcal{Q}_1^* before this departure. Inequality $0 \leq Q_1^*(t) - Q_1(t) \leq 2$ is valid until the time instant when both of these processes achieve a zero point at a moment of departure.

Case 2. ℓ is odd. In this case, the following two possible events can occur:

$$\sum_{k=1}^{\ell+1} \tau_k \leq \sum_{k=1}^{\ell+1} \chi_k,$$

or

$$\sum_{k=1}^{\ell+1} \tau_k > \sum_{k=1}^{\ell+1} \chi_k.$$

In the first situation $Q_1^*(\tau_{\ell+1}) = Q_1^*(\tau_\ell) = 2$ and $Q_1(\tau_{\ell+1}) = 1$, so $Q_1^*(\tau_{\ell+1}) - Q_1(\tau_{\ell+1}) = 1$, and after the time instant $\tau_{\ell+1}$ the difference $Q_1^*(t) - Q_1(t)$ will not be greater than 1 until the time instant when both of these processes achieve a zero point at a moment of departure.

In the second situation $Q_1^*(\tau_{\ell+1}) = 0$ and $Q_1(\tau_{\ell+1}) = 1$, so $Q_1^*(\tau_{\ell+1}) - Q_1(\tau_{\ell+1}) = -1$. After the time instant $\tau_{\ell+1}$ the difference $Q_1^*(t) - Q_1(t)$ will be not smaller than (-1) while both of these processes achieve a zero point at a moment of departure.

According to this construction, the processes $Q_1(t, \omega)$ and $Q_1^*(t, \omega)$ both are bounded (or unbounded) all together. Therefore, all of the earlier sample path arguments can be applied. \square

One of the possible generalizations of this construction is as follows.

Construction 3.5. *Let ξ_1, ξ_2, \dots be a sequence of independent and identically distributed random variables with mean $\frac{k}{\mu}$, where k is a positive integer. Consider a queueing system in which $A(t)$ is a point processes of arrivals, and $D(t)$ is a point process of departures. As earlier, let $t_i = \sum_{j=1}^i \tau_j$, $i = 1, 2, \dots$, be the moments of arrival, and let $x_i = \sum_{j=1}^i \chi_j$, $i = 1, 2, \dots$ be the corresponding time instants of departures, where the corresponding random variables τ_j and χ_j are generally dependent, but the new random variables ξ_1, ξ_2, \dots , which are defined via the random variables χ_1, χ_2, \dots by the representations:*

$$\xi_1 = \sum_{j=1}^k \chi_j, \quad \xi_2 = \sum_{j=k+1}^{2k} \chi_j, \quad \dots, \quad \xi_l = \sum_{j=(l-1)k+1}^{lk} \chi_j, \quad \dots \quad (3.7)$$

are independent and identically distributed as mentioned.

This construction is a generalization of that in Examples 1 and 2, where k was equal to 2 and, respectively, $\chi_{2n-1} + \chi_{2n}$ was equal to $\frac{2}{\mu}$, $n = 1, 2, \dots$. Under this construction Propositions 3.2 and 3.1 hold true as well, and technically more complicated sample-path proof is similar to that given in the above particular case.

The above construction can be extended more by different ways. One of the natural ways is to extend the representations of (3.7) as follows:

$$\xi_1 = \sum_{j=1}^{k_1} \chi_j, \quad \xi_2 = \sum_{j=k_1+1}^{k_1+k_2} \chi_j, \quad \dots, \quad \xi_l = \sum_{j=k_1+\dots+k_{l-1}+1}^{k_1+\dots+k_l} \chi_j, \quad \dots \quad (3.8)$$

where $k_1, k_2, \dots, k_l, \dots$ in (3.8) are independent and identically distributed positive integer random variables having an expectation. In further constructions, one can assume that these random variables are not identically distributed. In this case some additional assumptions are required as well.

Open problem. Find most general assumptions of generally dependent point processes $A(t)$ and $D(t)$, under which the queueing systems \mathcal{Q}_1 , \mathcal{Q}_2 and \mathcal{Q}_3 will be bounded or unbounded all together.

4. STABILITY OF JS-QUEUES

In this section we study the stability of JS-queues. We start from the simplest case of symmetric Σ_m queues. In Theorem 4.1 below, the assumption that the processes $A(t)$, $A'(t)$ and $D^{(j)}(t)$, $j = 1, 2, \dots, m$, all are mutually independent point processes is relaxed.

4.1. The case of symmetric queues. For Σ_m queues where $p_j = \frac{1}{m}$, $j = 1, 2, \dots, m$ instead of (2.1), (2.2) and (2.3) we assume

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{A(t-a) + A'(t-a) - D(t-a)}{t-a} = \lambda + \lambda' - \mu m \right\} = 1. \quad (4.1)$$

The processes $D^{(j)}(t)$, $j = 1, 2, \dots, m$ are assumed to be identically distributed. But the processes $A(t)$, $A'(t)$ and $D(t)$, according to the assumption (4.1), are generally dependent.

Theorem 4.1. *In addition to (4.1) assume that*

$$\lim_{a \rightarrow -\infty} \mathbf{P}\{A(t-a) + A'(t-a) - D(t-a) \in \mathcal{S}\} = 0 \quad (4.2)$$

for any bounded set $\mathcal{S} \in \mathbb{R}^1$. Then, the system Σ_m is stable if and only if the condition $\frac{\lambda}{m} + \frac{\lambda'}{m} < \mu$ is fulfilled.

Proof. Since the families $\{A^{(j)}(t)\}_{j \leq m}$ and $\{D^{(j)}(t)\}_{j \leq m}$ consist of identically distributed processes, then the family $\{A'^{(j)}(t)\}_{j \leq m}$ also consists of identically distributed processes, and according to (2.4) the family of the processes $\{Q^{(j)}(t)\}_{j \leq m}$ consists of identically distributed processes too. Observe that from (4.2), because the system Σ_m is symmetric, we also have

$$\lim_{a \rightarrow -\infty} \mathbf{P}\{A^{(j)}(t-a) + A'^{(j)}(t-a) - D^{(j)}(t-a) \in \mathcal{S}\} = 0 \quad (4.3)$$

for all $j = 1, 2, \dots, m$.

Notice, that if $\frac{\lambda}{m} + \frac{\lambda'}{m} < \mu$, then

$$\mathbf{P}\left\{\sup_{a \leq t < \infty} [A^{(j)}(t-a) + A'^{(j)}(t-a) - D^{(j)}(t-a)] < \infty\right\} = 1. \quad (4.4)$$

Indeed, according to (4.1), \mathbf{P} -a.s.

$$\lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a) + A'^{(j)}(t-a) - D^{(j)}(t-a)}{t-a} = \frac{\lambda}{m} + \frac{\lambda'}{m} - \mu < 0. \quad (4.5)$$

Hence, \mathbf{P} -a.s.

$$\lim_{a \rightarrow -\infty} [A^{(j)}(t-a) + A'^{(j)}(t-a) - D^{(j)}(t-a)] = -\infty,$$

and (4.4) follows from the fact that $A^{(j)}(t)$, $A'^{(j)}(t)$ and $D^{(j)}(t)$ all are càdlàg processes.

Next, taking into account (4.5) and assuming that $A^{(j)}(a) = A'^{(j)}(a) = D^{(j)}(a) = 0$, for the queue-length process $Q^{(j)}(t)$ one can write the following representation ($t \geq a$):

$$\begin{aligned} Q^{(j)}(t-a) &= [A^{(j)}(t-a) + A'^{(j)}(t-a) - D^{(j)}(t-a)] \\ &\quad - \inf_{a \leq s \leq t} [A^{(j)}(s-a) + A'^{(j)}(s-a) - D^{(j)}(s-a)]. \end{aligned} \quad (4.6)$$

Representation (4.6) is well-known (e.g. Borovkov [6], Whitt [49], p. 171) and is a consequence from the Skorokhod reflection principle (e.g. Kogan and Liptser [29] for typical application to queue-length processes).

Next, from (4.6) we have ($t \geq a$):

$$\begin{aligned} Q^{(j)}(t-a) &=_{st} \sup_{a \leq s \leq t} \{ [A^{(j)}(t-a) + A'^{(j)}(t-a) - D^{(j)}(t-a)] \\ &\quad - [A^{(j)}(s-a) + A'^{(j)}(s-a) - D^{(j)}(s-a)] \}, \end{aligned} \quad (4.7)$$

where $=_{st}$ means stochastic equivalence of the random variables on the left and right sides of (4.7), i.e. equality of one-dimensional distributions of the processes (which is enough for the purpose of the present paper).

From representation (4.7), relation (4.4) and the fact that the processes $A^{(j)}(t)$, $A'^{(j)}(t)$ and $D^{(j)}(t)$ all are càdlàg processes, it follows that there exists a bounded set \mathcal{S}_0 such that

$$\limsup_{a \rightarrow -\infty} \mathbf{P}\{Q^{(j)}(t-a) \in \mathcal{S}_0\} > 0,$$

and the sufficient condition is therefore proved. The necessary condition follows from the fact that (4.3) together with (4.4) imply the condition $\frac{\lambda}{m} + \frac{\lambda'}{m} < \mu$. \square

Remark 4.2. If $\frac{\lambda}{m} + \frac{\lambda'}{m} = \mu$, and the processes $A(t)$, $A'(t)$ and $D(t)$ all are non-trivial renewal processes, then we easily arrive at condition (4.2). However, there are examples where $\frac{\lambda}{m} + \frac{\lambda'}{m} = \mu$, but condition (4.2) is not fulfilled. Indeed, let $\tau_1 + \tau'_1 - \chi_1$ be a uniformly distributed random variable in $[-b, b]$, ($b > 0$), and let $\tau_{i+1} + \tau'_{i+1} - \chi_{i+1} = -(\tau_i + \tau'_i - \chi_i)$, $i \geq 1$. In this case (4.2) is not valid. Therefore condition (4.2) is meaningful.

4.2. The case of asymmetric queues. In following Theorem 4.3, our assumptions regarding the point processes $A(t)$, $D^{(j)}(t)$ are as follows. For dependent processes $A(t)$, $D^{(j)}(t)$ we suppose that the normalized processes $\frac{A(t-a)}{t-a}$ and $\frac{D^{(j)}(t-a)}{t-a}$ ($t > a$) converge, as $a \rightarrow -\infty$, to the corresponding limits λ and μ only in distribution, while

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{A(t-a) - D^{(j)}(t-a)}{t-a} = \lambda - \mu \right\} = 1. \quad (4.8)$$

Then, the process $A'(t)$ is assumed to satisfy the condition

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{A'(t-a)}{t-a} = \lambda' \right\} = 1. \quad (4.9)$$

Next Theorem 4.3 is related to the stability of \wp_m queues. Denote $\lambda_j = \lambda p_j$, $j^* = \arg \max_{1 \leq j \leq m} \lambda_j$ and $\Delta = \sum_{j=1}^m (\lambda_{j^*} - \lambda_j)$.

Theorem 4.3. *In addition to (4.8) and (4.9) assume that both*

$$\lim_{a \rightarrow -\infty} \mathbf{P}\{A(t-a) + A'(t-a) - D(t-a) \in \mathcal{S}\} = 0, \quad (4.10)$$

and

$$\lim_{a \rightarrow -\infty} \mathbf{P}\{A^{(j^*)}(t-a) - D^{(j^*)}(t-a) \in \mathcal{S}\} = 0 \quad (4.11)$$

for any bounded set \mathcal{S} . Then the system is stable if and only if one of the following two conditions is fulfilled:

$$\begin{cases} \lambda^{(j^*)} < \mu, & \text{if } \Delta \geq \lambda', \\ \lambda + \lambda' < m\mu, & \text{otherwise.} \end{cases}$$

Proof. The theorem can be proved by a slight modification of the earlier proof.

Suppose first that $\Delta \geq \lambda'$, and denote q_j the fraction of customer of the opportunistic traffic that being assigned to the j th queue. From the limiting relations

$$\begin{aligned} \lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a)}{t-a} &\stackrel{d}{=} \lambda_j, \\ \lim_{a \rightarrow -\infty} \frac{A'^{(j)}(t-a)}{t-a} &\stackrel{a.s.}{=} \lambda'_j, \end{aligned}$$

according to well-known Skorokhod's theorem [44], p. 281, one can conclude that there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$, with given there a family of processes $\left\{ \frac{A^{(j)}(t-a, \omega) + A'^{(j)}(t-a, \omega)}{t-a}, t > a \right\}$ such that for \mathbf{P} -almost all $\omega \in \Omega$,

$$\lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a, \omega) + A'^{(j)}(t-a, \omega)}{t-a} = \lambda_j + \lambda'_j.$$

Therefore, from the balance equations

$$\lambda_j + \lambda'_j = \lambda_j + \lambda'q_j = \varrho,$$

one can conclude that q_j^* must be equal to 0. Therefore

$$\lambda_{j^*} \geq \lambda_j + \lambda'_j$$

for all $j = 1, 2, \dots, m$, where $\lambda'_j = \lambda'q_j$. Then, in this probability space for \mathbf{P} -almost all $\omega \in \Omega$,

$$\lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a, \omega) + A'^{(j)}(t-a, \omega)}{t-a} \leq \lim_{a \rightarrow -\infty} \frac{A^{(j^*)}(t-a, \omega)}{t-a},$$

and, from (2.5) and coupling arguments of sample paths comparison

$$\lim_{a \rightarrow -\infty} \frac{Q^{(j)}(t-a, \omega)}{t-a} \leq \lim_{a \rightarrow -\infty} \frac{Q^{(j^*)}(t-a, \omega)}{t-a}.$$

Therefore, in the original probability space we have

$$\lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a) + A'^{(j)}(t-a)}{t-a} \stackrel{d}{\leq} \lim_{a \rightarrow -\infty} \frac{A^{(j^*)}(t-a)}{t-a},$$

and consequently,

$$\lim_{a \rightarrow -\infty} \frac{Q^{(j)}(t-a)}{t-a} \stackrel{d}{\leq} \lim_{a \rightarrow -\infty} \frac{Q^{(j^*)}(t-a)}{t-a}$$

for all $j = 1, 2, \dots, m$. Hence the problem reduces to the conditions of the stability of a single queueing system with an autonomous service mechanism, which is defined by the given arrival process $A^{(j^*)}(t)$ and departure process $D^{(j^*)}(t)$. Under assumptions (4.8) and (4.11) the necessary and sufficient condition of the stability is given by $\lambda_{j^*} < \mu$.

Let us now consider the opposite case $\Delta < \lambda'$ and assumption $\lambda + \lambda' < \mu m$. Then, there exist probabilities q_j , $j = 1, 2, \dots, m$, all are strictly positive, and satisfying the equality $\sum_{j=1}^m q_j = 1$. Under these probabilities, the opportunistic traffic is thinned into m processes such that almost surely

$$\lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a)}{t-a} = \lambda' q_j = \lambda'_j.$$

Indeed, since $\lambda + \lambda' < \mu m$, then there exists the value $\varrho = \frac{\lambda + \lambda'}{m} < \mu$ such that for all j

$$\varrho > \lambda_j,$$

and therefore,

$$q_j = \frac{\varrho - \lambda_j}{\lambda'} > 0. \quad (4.12)$$

Since $\lambda_j + q_j \lambda' = \varrho$, the family $\{A^{(j)}(t-a) + A'^{(j)}(t-a)\}_{j \leq m}$ consists of the processes having the same rate, i.e.

$$\lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a) + A'^{(j)}(t-a)}{t-a} \stackrel{d}{=} \varrho. \quad (4.13)$$

The possible values $\{q_j\}_{j \leq m}$ are unique, since otherwise, if there are different arrival rates $\lambda_j + \lambda' q'_j$, then one of the queues must be stochastically longer than other. Let j^* be the order number of the longer queue. Then q_{j^*} must be equal to 0, and we have a contradiction with (4.12).

According to (4.13) for each of m queue-length processes the arrival rate is the same. With the same arrival and departure intensities one can repeat the proof of Theorem 4.1 for each of queue-length processes. Therefore, $\varrho < \mu$ is the condition for the stability, and under condition (4.10) the system is stable if and only if $\varrho < \mu$. The theorem is proved. \square

4.3. Note on the relation to the Harris recurrent Markov processes. In this section we will not provide an exact proof of claims about interconnection of different definitions of stability. We only explain how the stability in the terms of the Harris recurrent Markov processes could be established in the particular case when the point processes $A(t)$, $A'(t)$ and $D^{(j)}(t)$, $j = 1, 2, \dots, m$ all are mutually independent renewal processes.

By the well-known standard methods of extending the phase space, the queue length process can be reduced to a Markov process. The Markov process in time t can be characterized by the queue-lengths in time t (the number of queues is m) and the residual times until the regeneration points in the renewal processes $A(t)$, $A'(t)$ and $D^{(j)}(t)$, $j = 1, 2, \dots, m$ in that time moment t . So, the dimension of the Markov process associated with the queueing system is $2m + 2$.

By weak regeneration points of this Markov process we mean the points at which the queues all become empty, i.e. the time instant when the m coordinates of the Markov process characterizing the queue-lengths all become equal to zero. (In any small neighborhood of this point there is a time instant when one of m queue-lengths processes is positive.)

Under the standard assumptions that $\mathbf{P}\{\tau_1 > N\} > 0$ and $\mathbf{P}\{\tau'_1 > N\} > 0$ for any positive N and any initial state, the empty system can be achieved with some positive probability α . Therefore, taking into account the stability of the system in the terms of Definition 2.1 (i.e. boundedness of the system in terms of this definition), one can conclude that the weak regeneration state can be accessed infinitely many times as duration of time increases to infinity.

So, the method of the proof applied to queueing systems with an autonomous service mechanism enables us to establish the Harris recurrence of the above Markov process easier than the direct method applied to the usual queueing systems. According to the sample path results obtained in the previous section, the Harris recurrence of Markov processes related to usual systems follows immediately from that of queueing system with an autonomous service mechanism.

5. LOAD-BALANCED NETWORKS AND THEIR STABILITY

5.1. Basic result. In the previous section, the necessary and sufficient conditions for the stability of systems φ_m have been established. In this section we extend above Theorem 4.3 for load-balanced networks associated with φ_m queueing systems. The main result of this section is Theorem 5.3 establishing the necessary and sufficient condition for the stability of load-balanced networks. Theorems 5.1 and 5.2 are preliminary results establishing only sufficient conditions for the stability.

The load-balanced network considered below is the following extension of the φ_m queueing system.

Assume that an arriving customer of the dedicated traffic occupies the server j with probability p_j ($j = 1, 2, \dots, m$), and $\sum_{j=1}^m p_j = 1$. After his service completion in the j th queue, a customer leaves the system with probability $1 - p_j^*$, remains at the same j th queue with probability $p_{j,j}$, goes to the different queue $i \neq j$ with probability $p_{j,i}$, and choose the shortest queue with probability $p_{j,sh}$, breaking ties at random. It is assumed that $p_j^* < 1$ at least for one of the indexes j . This model is called *load-balanced network*. The stability conditions of the Markovian variant of this network containing two stations only has been established by Kurkova [30].

Other variants of this network have also been studied in Dai, Hasenbein and Kim [15], Martin and Suhov [33], Vvedenskaya, Dobrushin and Karpelevich [47] and other papers.

Denote $\lambda_j = \lambda p_j$, $\Lambda_j = \lambda_j + \mu \sum_{i=1}^m p_{i,j}$, $j^* = \arg \max_{1 \leq j \leq m} \Lambda_j$.

We follow the assumptions given in Section 2 that the point processes $A(t)$, $A'(t)$ and renewal process $D(t)$ all are mutually independent. The assumption that $D^{(j)}(t)$, $j = 1, 2, \dots, m$, all are renewal processes can be then relaxed, so $D^{(j)}(t)$ can be special type mutually independent and identically distributed point processes which are described later in Section 3.5, but again they are assumed to be independent of the other processes $A(t)$ in this construction.

All the processes are assumed to be started at a and the assumption where $a \rightarrow -\infty$ is used.

The sufficient condition given by Theorem 5.1 is a straightforward extension of earlier Theorem 4.3 written now under simpler assumptions.

Theorem 5.1. *Assume that $\lambda_{j^*} \geq \Lambda_j$ for all $j \neq j^*$. Denote*

$$\Delta_1 = \sum_{j=1}^m (\Lambda_{j^*} - \lambda_j),$$

and

$$\Delta_2 = \sum_{j \neq j^*} (\lambda_{j^*} - \Lambda_j).$$

Then the load-balanced network is stable if one of the following two conditions is fulfilled:

$$\begin{cases} \Lambda_{j^*} < \mu, & \text{if } \Delta_2 \geq \lambda' + \mu \sum_{j=1}^m p_{j,sh}, \\ \lambda + \lambda' + \mu \left(\sum_{i=1}^m \sum_{j=1}^m p_{i,j} + \sum_{j=1}^m p_{j,sh} \right) < m\mu, & \text{if } \Delta_1 < \lambda'. \end{cases}$$

Proof. The proof of the theorem starts from the case $p_{j,sh} = 0$ for all $j = 1, 2, \dots, m$ and then the case $p_{j,sh} \geq 0$ for all $j = 1, 2, \dots, m$ is discussed.

Let us start from the case where $p_{j,sh} = 0$ for all $j = 1, 2, \dots, m$. Then, the system of equations for the queue-length processes started at a can be written as follows ($t > a$):

$$\begin{aligned} Q^{(j)}(t-a) = & A^{(j)}(t-a) + A'^{(j)}(t-a) + E^{(j)}(t-a) \\ & - \int_0^{t-a} \mathbf{1}_{\{Q^{(j)}(s-) > 0\}} dD^{(j)}(s), \end{aligned} \quad (5.1)$$

where the new process $E^{(j)}(t)$ given in (5.1) is a process, generated by *internal dedicated arrivals* to the j th queue. By internal dedicated arrivals to the j th queue we mean internal arrivals of the customers, who after their service completion in one or other queue are assigned to the j th queue. Recall that there is probability $p_{i,j}$ to be assigned from the queue i to the queue j . Relationship (5.1) is of the same type as that (2.5), and therefore all of the arguments of the earlier proof of Theorem 4.3 can be repeated. Specifically, in the case $\Delta_2 \geq \lambda'$ the system is stable if $\Lambda_{j^*} < \mu$.

In turn, the process $E^{(j)}(t)$ can be represented as $\sum_{i=1}^m E^{(i,j)}(t)$, where the point process $E^{(i,j)}(t)$ is generated by the customers who are assigned to the

j th queue after their service completion in the i th queue (in the case $i = j$ it is assumed that the customers decide to stay at the same queue). Notice, that

$$\lim_{a \rightarrow -\infty} \frac{E^{(i,j)}(t-a)}{t-a}$$

does exist with probability 1 (because the process $E^{(i,j)}(t)$ is generated by the procedure of thinning of the departure process $D^{(i)}(t)$), and

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{E^{(i,j)}(t-a)}{t-a} \leq \mu p_{i,j} \right\} = 1, \quad (5.2)$$

where the equality holds only in the case where the fraction of the i th queue idle period vanishes as $a \rightarrow -\infty$. Therefore under the condition $\Delta_1 < \lambda'$ the system is stable if $\lambda + \lambda' + \mu \sum_{i=1}^m \sum_{j=1}^m p_{i,j} < m\mu$.

Assume now that $p_{j,sh} \geq 0$ for all $j = 1, 2, \dots, m$. Then instead of (5.1) we have the equation

$$\begin{aligned} Q^{(j)}(t-a) &= A^{(j)}(t-a) + A'^{(j)}(t-a) + E_a^{(j)}(t) + E'_a{}^{(j)}(t) \\ &\quad - \int_0^{t-a} \mathbf{1}_{\{Q^{(j)}(s-) > 0\}} dD^{(j)}(s), \end{aligned} \quad (5.3)$$

where $E'^{(j)}(t)$ are the point processes associated with *internal opportunistic traffic* to the j th queue. By internal opportunistic traffic we mean the internal traffic of customers presenting in the queue, who after their service completion decide to join the shortest queue. In the case where the shortest queue is the j th queue, we say about opportunistic traffic to the j th queue. Again,

$$\lim_{a \rightarrow -\infty} \frac{E'^{(j)}(t-a)}{t-a}$$

does exist with probability 1 (because the process $E'^{(j)}(t-a)$ is generated by the procedure of thinning of the departure process $D^{(j)}(t-a)$), and similarly to (5.2) we have:

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{\sum_{j=1}^m E'^{(j)}(t-a)}{t-a} \leq \mu \sum_{j=1}^m p_{j,sh} \right\} = 1. \quad (5.4)$$

Therefore, the entire opportunistic traffic to the j th queue, being a sum of the processes $A^{(j)}(t - a)$ and $E^{(j)}(t - a)$, satisfies

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{\sum_{j=1}^m [A^{(j)}(t - a) + E^{(j)}(t - a)]}{t - a} \leq \lambda' + \mu \sum_{j=1}^m p_{j,sh} \right\} = 1. \quad (5.5)$$

Hence the proof of this theorem is similar to the proof of Theorem 4.3. In the case $\Delta_2 \geq \lambda' + \mu \sum_{j=1}^m p_{j,sh}$ the system is stable if $\Lambda_{j^*} < \mu$. In the other case $\Delta_1 < \lambda'$ the system is stable if $\lambda + \lambda' + \mu \left(\sum_{i=1}^m \sum_{j=1}^m p_{i,j} + \sum_{j=1}^m p_{j,sh} \right) < m\mu$. The conditions of the theorem are sufficient and not necessary, because the left-hand sides of (5.2), (5.4) and (5.5) contain the probability of inequalities, and the exact parameters of internal dedicated traffic as well as internal opportunistic traffic are unknown. \square

In order to formulate and prove a necessary and sufficient condition of the stability for the above load-balanced network, we first need to improve the sufficient condition given by Theorem 5.1. For this purpose, rewrite (5.2) as

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{E^{(i,j)}(t - a)}{t - a} = \varrho_i \mu p_{i,j} \right\} = 1,$$

where the value ϱ_i satisfies the inequality $0 < \varrho_i \leq 1$. The value ϱ_i is the fraction of time when the server of the i th queue is busy. Then the case of $\varrho_i = 1$ means that the server of the i th queue is busy almost always.

Let us consider the system of inequalities

$$\frac{\lambda_j + \mu \sum_{i=1}^m \varrho_i p_{i,j}}{\varrho_j^*} \leq \mu, \quad j = 1, 2, \dots, m. \quad (5.6)$$

The meaning of inequality (5.6) is the following. The left-hand side contains the total sum of rates of dedicated traffic to the j th queue divided to the traffic parameter ϱ_j^* of the j th queue. The total sum of rates of dedicated traffic of the j th queue consists of exogenous and internal arrivals to that j th queue, excluding the rates for joining the shortest queue customers. Since the rates associated with opportunistic traffic are excluded, there is the inequality ' \leq ' between the left and right sides. Thus, if the j th queue is never shortest, then the sum of the rates of the left-hand side divided to ϱ_j^* becomes equal to μ of the right-hand side. When the traffic parameter ϱ_j^* is greater than 1, the j th

queue increases to infinity with probability 1. Therefore, in the sequel we only consider the case when $\varrho_j^* \leq 1$ for all $j = 1, 2, \dots, m$. In this case $\varrho_j^* = \varrho_j$, and we therefore have

$$\lambda_j + \mu \sum_{i=1}^m \varrho_i p_{i,j} \leq \varrho_j \mu, \quad j = 1, 2, \dots, m. \quad (5.7)$$

Let us now write a so-called *balance equation*, taking into account also joining the shortest queue customers. We have

$$\lambda + \lambda' + \mu \sum_{j=1}^m \sum_{i=1}^m \varrho_i p_{i,j} = \mu \sum_{j=1}^m \varrho_j (1 - p_{j,sh}). \quad (5.8)$$

Now, we are ready to prove the *improved* sufficient condition for the stability. This version is also based on a straightforward extension of Theorem 4.3.

Theorem 5.2. *The load-balanced network is stable if there exists*

$$\varrho^* = \max_{1 \leq j \leq m} \varrho_j,$$

satisfying the condition $\varrho^ < 1$, where the values ϱ_j , $j = 1, 2, \dots, m$, are defined by (5.7) and (5.8).*

Proof. Let $\Lambda_j = \lambda_j + \mu \sum_{i=1}^m \varrho_i p_{i,j}$, let $j^* = \arg \max_{1 \leq j \leq m} \Lambda_j$, and let $\Delta = \sum_{j=1}^m (\Lambda_{j^*} - \Lambda_j)$. In the case $\Delta > \lambda' + \mu \sum_{j=1}^m \varrho_j p_{j,sh}$ the j^* -th queue is never shortest, and therefore, following the proof of Theorem 4.3, the system is stable if $\Lambda_{j^*} < \mu$. Therefore from (5.6) we have $\varrho_{j^*} < 1$, and because the j^* -th queue is the longest queue, we have $\varrho^* = \varrho_{j^*} \geq \varrho_j$, $j = 1, 2, \dots, m$. Therefore $\varrho_j < 1$ for all $j = 1, 2, \dots, m$ is a sufficient condition of stability for this case.

Let us now consider the opposite case, where $\Delta \leq \lambda' + \mu \sum_{j=1}^m \varrho_j p_{j,sh}$. As in the proof of Theorem 4.3, in this case the arrival rate to all m queues is the same, and therefore $\varrho_1 = \varrho_2 = \dots = \varrho_m$. Thus, the only two cases are there as $\varrho_j < 1$ or $\varrho_j = 1$ for all $j = 1, 2, \dots, m$. In the case $\varrho_j < 1$, the stability result is analogous to that of Theorem 4.3, since in this case from (5.8) we obtain

$$\lambda + \lambda' + \mu \sum_{j=1}^m \sum_{i=1}^m p_{i,j} + \mu \sum_{j=1}^m p_{j,sh} < m\mu.$$

The theorem is proved. □

Now in order to formulate and prove a necessary and sufficient condition for the stability, let us consider the following linear programming problem in \mathbb{R}^{m+1} :

$$\text{Minimize } x_{m+1} \quad (5.9)$$

subject to the restrictions:

$$\lambda_j + \mu \sum_{i=1}^m x_i p_{i,j} \leq x_j \mu, \quad j = 1, 2, \dots, m, \quad (5.10)$$

$$\lambda + \lambda' + \mu \sum_{j=1}^m \sum_{i=1}^m x_i p_{i,j} = \mu \sum_{j=1}^m x_j (1 - p_{j,sh}), \quad (5.11)$$

$$x_j \leq x_{m+1}, \quad j = 1, 2, \dots, m. \quad (5.12)$$

Observe, that the restrictions (5.10) and (5.11) correspond to (5.7) and (5.8), where the values ϱ_j are replaced with unknown x_j . The functional of (5.9) and inequalities (5.12) are associated with the condition of Theorem 5.2: $\max_{1 \leq j \leq m} \varrho_j < 1$. x_{m+1} is an additional variable; thus the linear programming (5.9)-(5.12) is a mini-max problem. That is, if the minimum of x_{m+1} is achieved in some point $x_{m+1}^* < 1$, then all of the components of the vector $(x_1^*, x_2^*, \dots, x_{m+1}^*)$ associated with this solution are less than 1, and there exists a solution of the system (5.7) and (5.8) with $\varrho_j < 1$ for all $j = 1, 2, \dots, m$. Therefore in the following the vector associated with a solution of the problem (5.9)-(5.12) is denoted $(\varrho_1, \varrho_2, \dots, \varrho_m)$. Otherwise if $x_{m+1}^* \geq 1$, then we set $\varrho_j = 1, j = 1, 2, \dots, m$.

Next, denote by $A^{(j)}(t) + E^{(j)}(t)$ the dedicated arrival process. Its relation to the initial processes $A(t)$ and $D^{(j)}(t)$ is as follows. Each arrival of the initial process $A(t)$ is forwarded to the queue j with probability p_j , and each customer served in the queue i returns to the j th queue with probability $p_{i,j}$. Then, the process $A^{(j)}(t) + E^{(j)}(t)$ is a sum of all arrivals of external and internal dedicated traffics, and

$$\mathbf{P} \left\{ \lim_{a \rightarrow -\infty} \frac{A^{(j)}(t-a) + E^{(j)}(t-a)}{t-a} = \Lambda_j = \lambda_j + \mu \sum_{i=1}^m \varrho_i p_{i,j} \right\} = 1,$$

where $\varrho_j, j = 1, 2, \dots, m$, are a solution of the linear programming given by (5.9)-(5.12). Now let $j^* = \arg \max_{1 \leq j \leq m} \Lambda_j$. We have the following theorem.

Theorem 5.3. *Assume that the both*

$$\lim_{a \rightarrow -\infty} \mathbf{P}\{A^{(j^*)}(t-a) + E^{(j^*)}(t-a) - D^{(j^*)}(t-a) \in \mathcal{S}\} = 0, \quad (5.13)$$

and

$$\lim_{a \rightarrow -\infty} \mathbf{P}\{A(t-a) + A'(t-a) + E(t-a) + E'(t-a) - D(t-a) \in \mathcal{S}\} = 0, \quad (5.14)$$

for any bounded set \mathcal{S} , where $E(t) = \sum_{j=1}^m E^{(j)}(t)$ is the point process associated with all internal arrivals of dedicated traffic, $E'(t)$ is the point process associated with all internal arrivals of opportunistic traffic, $D(t) = \sum_{j=1}^m D^{(j)}(t)$. Then the load-balanced network is stable if and only if $\max_{1 \leq j \leq m} \varrho_j < 1$.

Remark 5.4. Conditions (5.13) and (5.14) are verifiable conditions. As soon as the linear programming problem is solved and we know the vector of solution $(\varrho_1, \varrho_2, \dots, \varrho_m)$, the unknown processes $E(t)$ and $E'(t)$ as well as $E^{(j)}(t)$ and $E'^{(j)}(t)$ can be easily modelled via derivative processes $D^{(j)}(t)$ ($j = 1, 2, \dots, m$). Note also, that above conditions (5.13) and (5.14) are automatically fulfilled if $\max_{1 \leq j \leq n} \varrho_j < 1$.

5.2. Note on the relation to the Harris recurrent Markov processes.

Let us now explain the relation to the stability in the terms of the Harris recurrent Markov processes in the case of load-balanced networks in the particular case when the point processes $A(t)$, $A'(t)$ and $D^{(j)}(t)$, $j = 1, 2, \dots, m$ all are renewal processes. In this case the additional processes $E^{(i,j)}(t)$ and $E'^{(j)}(t)$ ($i, j = 1, 2, \dots, m$) all are renewal processes. By extending the phase space the queue-length process of the load-balanced network can be reduced to a Markov process, which is characterized by the queue-lengths in time t and the residual times until regeneration points in all of renewal processes such as $A(t)$, $A'(t)$ and $D^{(j)}(t)$, $E^{(i,j)}(t)$ and $E'^{(j)}(t)$ ($i, j = 1, 2, \dots, m$). The dimension of the Markov process in this case is $m^2 + 3m + 2$. (In the m parallel queues it was $2m + 2$. We add to this number m^2 generated by the set of the residual times of the processes $E^{(i,j)}(t)$ and m generated by the set of residual times generated by the processes $E'^{(j)}(t)$ and we obtained the desired value $m^2 + 3m + 2$.)

As in the case of usual parallel queues, by weak regeneration points of this Markov process we mean the points at which the queues all become empty.)

The ‘standard assumptions’ in this case are as follows: $\mathbf{P}\{\tau_1 > N\} > 0$, $\mathbf{P}\{\tau'_1 > N\} > 0$ and $\mathbf{P}\{\chi_1^{(1)} > N\} > 0$ for any positive N . We use the fact that $\chi_i^{(j)}$, $i = 1, 2, \dots$; $j = 1, 2, \dots, m$ all are identically distributed, and the processes $E^{(i,j)}(t)$ and $E'^{(j)}(t)$ ($i, j = 1, 2, \dots, m$) all are generated by the independent and identically distributed processes $D^{(j)}(t)$.

The further arguments are the same as in the case of parallel queues considered in Section 4.3. Again, we use the boundedness of all queue-length processes in terms of Definition 2.1. According to this boundedness, the weak regeneration points will be accessed infinitely many times as the duration of time increases to infinity.

6. CONCLUDING REMARKS

In this paper we established the stability of different type joint-the-shortest-queue models including load-balanced networks. The statements of stability are established under quite general assumptions on arrival and departure processes by reduction to the corresponding models with an autonomous service mechanism.

Now we discuss how these results can be extended to the models of queues and networks allowing batch arrivals and batch departures. For this purpose, consider the queueing system with batch arrivals and departures and autonomous service. For this queueing system let $\mathcal{A}(t)$ denote arrival process and let $\mathcal{D}(t)$ departure process, both marked point processes. (All of the processes considered in this section are assumed to start at zero.) For the sake of simplicity suppose that the marks of the point process $\mathcal{D}(t)$ all are of the constant size c (c is a positive integer number), and therefore $\mathcal{D}(t) = cD(t)$. Then, the queue-length process $Q(t)$ has the following representation (see [3]):

$$Q(t) = \mathcal{A}(t) - \sum_{i=1}^c \int_0^t \mathbf{1}_{\{Q(s-) \geq i\}} dD(s).$$

It was shown in [3] that by using Skorokhod's reflection principle we arrive at the equation

$$Q(t) = [\mathcal{A}(t) - \mathcal{D}(t)] - \inf_{s \leq t} [\mathcal{A}(s) - \mathcal{D}(s)], \quad (6.1)$$

which is similar to that of the process with ordinary departures. The assumption, that the marks of departure process are a constant c , is specific and associated with concrete models considered in [3]. Representation (6.1) remains in force in general, when a departure process is an arbitrary marked point process with mutually independent identically distributed marks. Representation (6.1) is easily generalized to the case of JS-queue models. Specifically, for the queue-length process in the j th server of the model φ_m we have the similar equation

$$\begin{aligned} Q^{(j)}(t) =_{st} & [\mathcal{A}^{(j)}(t) + \mathcal{A}'^{(j)}(t) - \mathcal{D}^{(j)}(t)] \\ & - \inf_{s \leq t} [\mathcal{A}^{(j)}(s) + \mathcal{A}'^{(j)}(t) - \mathcal{D}^{(j)}(s)], \end{aligned} \quad (6.2)$$

where $\mathcal{A}'^{(j)}(t)$ is the corresponding notation for an opportunistic traffic to the j th server of the JS-queue model (see ref. (4.6) for comparison). Thus, the case of batch arrivals and departures is a direct extension of the case of ordinary arrivals and departures, and the conditions for stability are similar.

ACKNOWLEDGEMENTS

The advice of Professors Rafael Hassin, Robert Liptser, Yuri Suhov, Gideon Weiss and Ward Whitt helped very much to substantially improve the presentation. The research was supported by Australian Research Council, grant #DP0771338.

REFERENCES

- [1] ABRAMOV, V.M. (2000). A large closed queueing network with autonomous service and bottleneck. *Queueing Systems*, **35**, 23-54.
- [2] ABRAMOV, V.M. (2004). A large closed queueing network containing two types of node and multiple customers classes: One bottleneck station. *Queueing Systems*, **48**, 45-73.
- [3] ABRAMOV, V.M. (2008). The effective bandwidth problem revisited. *Stoch. Models*, **24**, 527-557.
- [4] ABRAMOV, V.M. (2008). Large closed queueing networks in semi-Markov environment and their applications. *Acta Appl. Math.*, **100**, 201-226.

- [5] BACCELLI, F. AND FOSS, S. (1994). Stability of Jackson-type queueing networks. *Queueing Systems*, **17**, 5-72.
- [6] BOROVKOV, A.A. (1976). *Stochastic Processes in Queueing Theory*. Springer-Verlag, Berlin.
- [7] BOROVKOV, A.A. (1984). *Asymptotic Methods in Queueing Theory*. John Wiley, New York.
- [8] BOROVKOV, A.A. (1986). Limit theorems for queueing networks. I. *Theor. Prob. Appl.* **31**, 413-427.
- [9] BOROVKOV, A.A. (1987). Limit theorems for queueing networks. II. *Theor. Prob. Appl.* **32**, 257-272.
- [10] BOROVKOV, A.A. (1998). *Ergodicity and Stability of Stochastic Processes*. John Wiley, New York.
- [11] BRAMSON, M.D. (2008). Stability of queueing networks. *Probab. Surveys*, **5**, 169-345.
- [12] BRAMSON, M.D. (2008). *Stability of queueing networks*. Springer-Verlag, Heidelberg.
- [13] BRANDT, A., FRANKEN, P. AND LIZEK, B. (1990). *Stationary Stochastic Models*, Akademie-Verlag/Wiley, Berlin/Chichester.
- [14] DAI, J.G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Ann. Appl. Probab.*, **5**, 49-77.
- [15] DAI, J.G., HASENBEIN, J.J. AND KIM, B. (2007). Stability of join-the-shortest-queue networks. *Queueing Syst.* **57**, 129-145.
- [16] EL-TAHA, M. AND STIDHAM, S. (1999). *Sample-Path Analysis of Queueing Systems*, Kluwer, Dordrecht.
- [17] FLATTO, L., AND MCKEAN, H.P. (1977). Two queues in parallel. *Comm. Pure Appl. Math.* **30**, 255-263.
- [18] FOLEY, R.D. AND McDONALD, R.D. (2001). Join-the-shortest-queue: Stability and exact asymptotics. *Ann. Appl. Probab.* **11**, 569-607.
- [19] FOSS, S., AND CHERNOVA, N. (1991). On the ergodicity of multichannel not fully accessible communication systems. *Problems of Information Transmission*, **27**, 94-99.
- [20] FOSS, S., AND CHERNOVA, N. (1998). On the stability of partially accessible queue with state-dependent routing. *Queueing Systems*, **29**, 55-73.
- [21] FRICKER, C. (1986). Etude d'une file GI/G/1 á service autonome (avec vacances du serveur). *Adv. Appl. Probab.*, **18**, 283-286.
- [22] FRICKER, C. (1987). Note sur un modele de file GI/G/1 á service autonome (avec vacances du serveur). *Adv. Appl. Probab.*, **19**, 289-291.
- [23] GELENBE, E. AND IASNOGORODSKI, R. (1980). A queue with server of walking type (autonomous service). *Ann. Inst. H. Poincare*, **16**, 63-73.
- [24] HALFIN, S. (1985). The shortest queue problem. *J. Appl. Probab.* **22**, 865-878.

- [25] IGLEHART, D. AND WHITT, W. (1970). Multiple channel queues in heavy traffic. I. *Adv. Appl. Prob.*, **2**, 150-177.
- [26] KASPI, H. AND MANDELBAUM, A. (1992). Regenerative closed queueing networks. *Stoch. Stoch. Rep.*, **39**, 239-258.
- [27] KASPI, H. AND MANDELBAUM, A. (1994). On Harris recurrence in continuous time. *Math. Operat. Res.*, **19**, 211-222.
- [28] KEILSON, J. (1979). *Markov Chain Models. Rarity and Exponentiality*. Springer, Heidelberg.
- [29] KOGAN, YA., AND LIPTSER, R.SH. (1993). Limit non-stationary behavior of large closed queueing network with bottlenecks. *Queueing Systems*, **14**, 33-55.
- [30] KURKOVA, I.A. (2001). A load-balanced network with two servers. *Queueing Systems*, **37**, 379-389.
- [31] KURKOVA, I.A., AND SUHOV, YU.M. (2003). Malyshev's theory and JS-queues. Asymptotics of stationary probabilities. *Ann. Appl. Probab.* **13**, 1313-1354.
- [32] LOYNES, R. (1962). The stability of queues with non-independent interarrival and service times. *Proc. Camb. Phil. Soc.* **58**, 497-520.
- [33] MARTIN, J.B., AND SUHOV, YU.M. (1999). Fast Jackson networks. *Ann. Appl. Probab.* **9**, 854-870.
- [34] MEYN, S.P. AND DOWN, D. (1994). Stability of generalized Jackson networks. *Ann. Appl. Prob.* **4**, 124-148.
- [35] MEYN, S.P. AND TWEEDIE, R.L. (1992). Stability of Markov processes. I. Criteria for discrete time chains. *Adv. Appl. Prob.*, **24**, 542-574.
- [36] MEYN, S.P. AND TWEEDIE, R.L. (1993). Stability of Markov processes. II. Continuous time processes and sampled paths. *Adv. Appl. Prob.*, **25**, 487-517.
- [37] MEYN, S.P. AND TWEEDIE, R.L. (1993). Stability of Markov processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. Appl. Prob.*, **25**, 518-548.
- [38] MEYN, S.P. AND TWEEDIE, R.L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, Berlin.
- [39] MITZENMACHER, M.D. (1996). The power of two choices in randomized load balancing. PhD thesis. University of California at Berkeley.
- [40] OREY, S. (1971). *Lecture Notes on Limit Theorems for Markov Chains Transition Probabilities*. Van Nostrand Reinhold Co., London.
- [41] REVUZ, D. (1975). *Markov Chains*. North-Holland Publishing Co., Amsterdam.
- [42] SHARIFNIA, A. (1997). Instability of the join-the-shortest-queue and FCFS policies in queueing systems and their stabilization. *Operations Research*, **45**, 309-314.
- [43] SIGMAN, K. (1990). The stability of open queueing networks. *Stoch. Proces. Appl.* **35**, 11-25.

- [44] SKOROKHOD, A.V. (1956). Limit theorems for stochastic processes. *Theor. Prob. Appl.*, **1**, 261-290.
- [45] SUHOV, YU.M., AND VVEDENSKAYA, N.D. (2002). Fast Jackson networks with dynamic routing. *Probl. Inform. Transmission*, **38**, 136-159.
- [46] TANDRA, R., HEMACHANDRA, N. AND MANJUNATH, D. (2004). Job minimum cost queue for multiclass customers. Stability and Performance bounds. *Probability in the Engineering and Informational Sciences*, **18**, 445-472.
- [47] VVEDENSKAYA, N.D., DOBRUSHIN, R.L., AND KARPELEVICH, F.I. (1996). A queueing system with selection the shortest of two queues: An asymptotic approach. *Probl. Inform. Transmission*, **32**, 15-27.
- [48] VVEDENSKAYA, N.D. AND SUHOV, YU.M. (2004). Functional equations in asymptotic problems of queueing theory. *Journal of Mathematical Sciences (N.Y.)*, **120**, 1255-1276.
- [49] WHITT, W. (2002). *Stochastic Process Limits*. Springer-Verlag, Heidelberg.
- [50] WINSTON, W. (1977). Optimality of the shortest line discipline. *J. Appl. Probab.* **14**, 181-189.

SCHOOL OF MATHEMATICAL SCIENCES, MONASH UNIVERSITY, CLAYTON CAMPUS, BUILDING 28, LEVEL 4, WELLINGTON ROAD, CLAYTON, VICTORIA 3800, AUSTRALIA

E-mail address: Vyacheslav.Abramov@sci.monash.edu.au