DIAS-STP-00-05

FAST JACKSON-TYPE NETWORKS WITH DYNAMIC ROUTING

Yu.M. Suhov$^{1,2)}$ and N.D. Vvedenskaya$^{1)}$

AMS 1991 subject classifications: Primary 90B15; Secondary 60K25, 60J27.

Key words and phrases: Jackson network, queueing network, dynamic routing, Markov process, system of ODEs, fixed point, super-exponential decay.

Abstract. We propose a new class of models of queueing networks with load-balanced dynamic routing. The paper extends earlier works, including [FC], [FMcD], [VDK], where systems with no feedback were considered. The main results are: (a) a sufficient condition for positive recurrence of the arising Markov process and (b) a limiting mean-field picture where the process becomes deterministic and is described by a system of non-linear ODEs.

0. Introduction

Historic background. This paper proposes a class of queueing network models with dynamic routing based on the principle of balanced load. The idea of dynamic routing is to select a path across a network in such a way that it minimises (i) the delivery time (or the end-to-end delay) of a given task, and (ii) the occupancy of buffers in the network. These goals do not always agree; besides, a decision is to be taken on the basis of a limited

[1)] Institute for Information Transmission Problems, Russian AS, GSP-4 Moscow, 101447, Russia E-mail: ndv@iitp.ru

[2)] Statistical Laboratory, DPMMS, University of Cambridge, Cambridge CB2 1SB, UK, and St John's College, Cambridge CB2 1TP, UK E-mail: yms@statslab.cam.ac.uk

amount of available information. To our knowledge, until recently, in the literature there was no mathematical model proposed, of a queueing network with dynamical routing, which could be studied rigorously[1,2]; the first example of such a model was a queueing system introduced in [VDK] (and independently, but somewhat later in [M]). See also [VS] and [T]. The original model was then modified in [MS] to include a class of Jackson-type networks, but the dynamic routing principle was still reduced in [MS] to the choice between servers from a given station. We refer the reader to the introductory section of [MS] for a discussion of the approach adopted in the above papers; a review of the available literature can also be found in [KPS].

The principle of dynamic routing proposed in the above papers is to select a server with the shortest queue among a sample collection of servers chosen at random. E.g., in the model considered in [VDK] there are N identical exponential servers, each with an infinite buffer and service rate one. The exogenous flow is Poisson with rate $N\lambda$; service times and arrival times are all independent. Upon arrival, each task chooses m servers at random ($m > 1$), and joins the one whose queue is the shortest. If $\lambda < 1$, the system is described by a positive recurrent Markov process whose state is represented by a *tail histogram* identifying, for each $n = 0, 1, 2, \ldots$, a proportion $r(n)$ of servers with at least n tasks in the queue. This Markov process has a unique invariant distribution π_N, and the main result of [VDK] is that the expected value $\mathbb{E}_{\pi_N} r(n)$ of proportion $r(n)$ converges, as $N \to \infty$, to $\lambda^{(m^n-1)/(m-1)}$ which gives a super-exponential decay as $n \to \infty$. This result contrasts with a model where each task selects a server independently and completely at random (which corresponds to the previous scheme with $m = 1$): here, $\mathbb{E}_{\pi_N} r(n) = \lambda^n$ (a geometric, or exponential decay).

Similarly, in the model considered in [MS] there is a set of stations $\mathcal{J} = \{1, \ldots, J\}$, station i containing N identical exponential servers, each with an

[1] See [FIM] for references to works on two-server systems. However, these do not include networks with feedback.

[2] Loss networks with dynamic routing have been discussed in [Ke]. However, loss networks do not pose such challenging problems as the existence of a stationary distribution.

infinite buffer and service rate μ_i; for simplicity, assume that $\mu_i = 1$. The exogenous flow to station i is Poisson with rate $N\lambda_i$; all service times are mutually independent and independent of the arrival times. It is convenient to introduce the vector $\overline{\lambda} = (\lambda_j, \, j \in \mathcal{J})$. Upon arrival in station i, each task chooses m servers at random, and joins the one whose queue is the shortest. After completing service in station i, the task is dispatched to station j with probability $P_{i,j}$ and exits the network with probability $1 - \sum_j P_{i,j}$. Here $P = (P_{i,j})$ is a (sub-stochastic) Jackson routing $(J \times J)$ matrix; one assumes that matrix $I - P$ is invertible. If the vector $\overline{\rho} = \overline{\lambda}(I-P)^{-1}$ with components $\rho_1, ..., \rho_J$ obeys $\rho_i < 1$, $i \in \mathcal{J}$, the network is described by a positive recurrent Markov process whose state is now represented by a vector tail histogram identifying, for each $i \in \mathcal{J}$ and $n = 0, 1, 2, ...$, a proportion $r_i(n)$ of servers in station i with at least n tasks in the queue. As before, this Markov process has a unique invariant distribution π_N; the main result of [MS] is that the expected value $\mathbb{E}_{\pi_N} r_i(n)$ converges, as $N \to \infty$, to $\rho_i^{(m^n-1)/(m-1)}$. This contrasts with the corresponding Jackson model where, upon arrival at a station, the task joins a randomly chosen queue: here the expectation $\mathbb{E}_{\pi_N} r_i(n) = \rho_i^n$.

On the other hand, [FC] and [FMcD] deal with a system of J stations, each containing a single exponential server of rate 1. The arrival is described by a collection of independent Poisson flows $\Xi_{\mathcal{K}}$ of rates $\lambda_{\mathcal{K}} \geq 0$, for each non-empty subset $\mathcal{K} \subseteq \mathcal{J}$, and the rule is that the tasks from $\Xi_{\mathcal{K}}$ joins the station from $\mathcal{K}$ with the shortest queue. Upon completing the service, the task leaves the system. The condition of positive recurrence here is that $\sum_{\mathcal{K} \subset \mathcal{A}} \lambda_{\mathcal{K}} < \# \mathcal{A}$ for any non-empty $\mathcal{A} \subseteq \mathcal{J}$. There is no explicit formulas for the equilibrium distribution, but some tail asymptotics are available, at least for $J = 2$.

Description of the model under consideration. In this paper we deal with a generalisation of the models discussed in [FC], [FMcD] and [MS]. Namely, comparing with [MS], we allow a task to choose a server from a sample that is *not* confined to a single station, and comparing with [FC], [FMcD], we allow a Jackson-type feedback, or networking. The collection of stations is $\mathcal{J} = \{1, \dots, J\}$, each station containing N single exponential servers of rate one. The network model is determined by the exogenous Poisson flows $\Xi_{\mathfrak{m}}$ of rates $N\lambda_{\mathfrak{m}}$ and routing probabilities p_j^*, $p_{j,\mathfrak{m}} \in [0, 1]$, $j \in \mathcal{J}$. Here, $\mathfrak{m}$ is a positive integer mass distribution on $\mathcal{J}$, i.e., a function $i \in \mathcal{J} \mapsto m_i \in \mathbb{Z}_+$

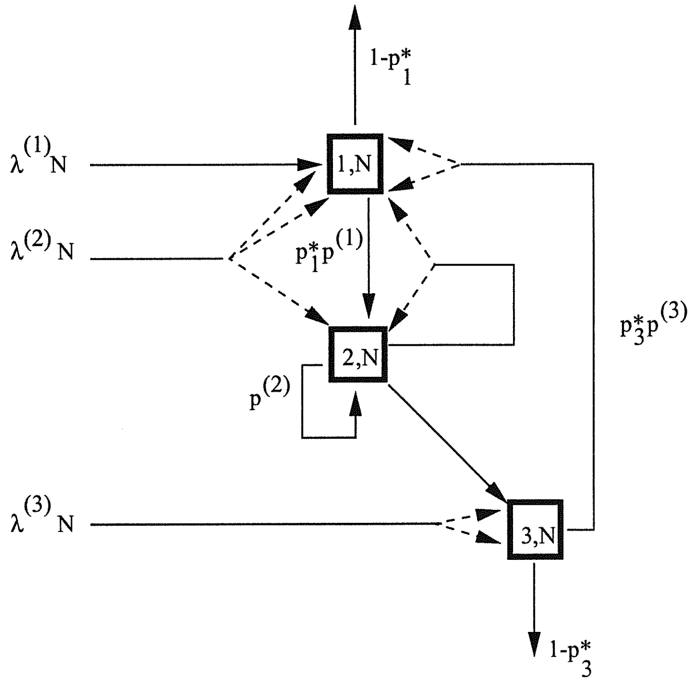

Figure 1. The scheme of the network model

$= \{0, 1, \dots\}$ with $|\mathfrak{m}| = \sum_{i \in \mathcal{J}} m_i > 0$. It describes how many inspections a given task performs in a given station; for this reason we call it an inspection number distribution. For simplicity, we assume that $\lambda_{\mathfrak{m}} > 0$ and $p_{j,\mathfrak{m}} > 0$ for a finitely many $\mathfrak{m}$ only, and $\sum_{\mathfrak{m}} p_{j,\mathfrak{m}} = 1 \; \forall \; j \in \mathcal{J}$. The exogenous Poisson flows are assumed independent, and a task arrived in flow $\Xi_{\mathfrak{m}}$ chooses at random m_i servers in each station $i \in \mathcal{J}$ (the choice is with replacement, so some servers may be chosen repeatedly) and then joins the shortest queue from the sample. After completing the service in station j, the task quits or remains in the network with probabilities $1 - p_j^*$ and p_j^*; in the latter case it picks up a positive inspection number distribution $\mathfrak{m}$ with probability $p_{j,\mathfrak{m}}$ and then again chooses at random m_i servers in each station $i \in \mathcal{J}$ and joins the shortest queue from the sample. When occur, the ties are broken at random (i.e., if $m^* \leq |\mathfrak{m}|$ servers from the sample have the shortest queue-length, any of them can be chosen with probability $1/m^*$).

The above model is called shortly an L-model or L-network (for load-balancing); it is determined by the parameters J, N, $\lambda_{\mathfrak{m}}$, p_j^* and $p_{j,\mathfrak{m}}$. A notable feature is an additional symmetry between N servers within a station. Also, for N large, the principle of dynamic routing is somehow 'softened': the task does not inspect all queues, but only a portion.

Some of our results for the L-model hold for any $N = 1, 2, \ldots$. Others are established in the limit $N \to \infty$. Under a simplifying assumption that $p_{j,\mathfrak{m}}$ does not depend on j, or even equals $\lambda_{\mathfrak{m}}/\Lambda$, where $\Lambda = \sum_{\tilde{\mathfrak{m}}} \lambda_{\tilde{\mathfrak{m}}}$, we can produce a more detailed information.

A an example of an L-network is presented on Figure 1. In this example, there are three exogenous flows, of rates $\lambda^{(i)} N = \lambda_{\mathfrak{m}^{(i)}} N$, with three inspection number distributions $\mathfrak{m}^{(i)} = \{m_j^{(i)},\ i, j = 1, 2, 3\}$, as follows : a) $m_2^{(1)} = m_3^{(1)} = 0$, $m_1^{(1)} = 1$ (a single inspection in station 1), b) $m_3^{(2)} = 0$, $m_2^{(2)} = 1$, $m_1^{(2)} = 2$ (a double inspection in station 1 and a single in station 2), c) $m_3^{(3)} = 2$, $m_2^{(2)} = m_1^{(2)} = 0$ (a double inspection in station 3). After service in station 1 or 3, the tasks quit with probabilities $1 - p_1^*$ and $1 - p_3^*$, while after service in station 2, they always return to the network: $p_2^* = 1$. Finally, the featured transition probabilities $p^{(k)} = p_{k,\bar{\mathfrak{m}}_k}$, after completing service in station $k = 1, 2, 3$, correspond to a) $\bar{m}_1^{(1)} = \bar{m}_3^{(1)} = 0$, $\bar{m}_2^{(1)} = 1$, b) $\bar{m}_1^{(2)} = \bar{m}_3^{(1)} = 0$, $\bar{m}_2^{(2)} = 1$ (a single inspection in station 2), c) $\bar{m}_1^{(3)} = 2$, $\bar{m}_3^{(1)} = \bar{m}_2^{(2)} = 0$ (a double inspection in station 1). There may also be other transitions (omitted in order not to overload the diagram).

Our exposition is carried in Sections 1–7 below. Sections 1 and 2 present results at an informal level. In Section 1 we discuss the capacity domain of an L-network and in Section 2 properties of the limit $N \to \infty$. In Section 3 a formal mathematical background is provided and the main theorems stated. The rest of the paper is devoted to the proofs. In Section 4 we prove Theorem 1. Section 5 is devoted to the analysis of the limiting system (2.1), (2.2) and its stationary version (2.3). In Section 6 we establish the convergence to the limiting picture as $N \to \infty$. Finally, in Section 7 we analyse a particular case where probabilities p_j^* and $p_{j,\mathfrak{m}}$ do not depend on j, the station where previous service has been completed.

1. The capacity domain

General bounds. The first question that arises is when the L-network is in a sub-critical regime, i.e., has a proper equilibrium, or stationary, distribution. More precisely, the above model leads to a denumerable continuous-time Markov process whose state is represented by a collection the queue lengths $Q_{j,s}(t)$, where $j \in \mathcal{J}$ labels the stations and $s = 1, \dots, N$ the servers (within a given station). We call it the q-process. Most of the time we will work with a 'symmetrised' Markov process, called the r-process, whose state is given by a (vector) tail histograms $r_j(n)$, $j \in \mathcal{J}$, $n \in \mathbb{Z}_+$, where $r_j(n)$ is the proportion of the servers in station j with the queue length $\geq n$. We propose a sufficient (but not necessary) condition for these processes to be positive recurrent, and hence to have a unique equilibrium distribution.

Condition 1: $\forall$ subset of stations $\mathcal{K} \subseteq \mathcal{J}$,

$$\zeta_{\mathcal{K}} < \#\mathcal{K}, \tag{1.1.1}$$

or, equivalently,

$$\nu < 1. \tag{1.1.2}$$

Here,

$$\nu = \max_{\mathcal{K} \subseteq \mathcal{J}} \frac{1}{\#\mathcal{K}} \zeta_{\mathcal{K}} < 1. \tag{1.2.1}$$

and

$$\zeta_{\mathcal{K}} = \sum_{\mathrm{m}:\, \mathrm{m}\lceil \mathcal{K}^{\mathrm{c}} = 0} \left(\lambda_{\mathrm{m}} + \sum_{j \in \mathcal{J}} p_j^* p_{j,\mathrm{m}} \right) . \tag{1.2.2}$$

Condition 1 describes the sub-criticality domain in a *majorising* system, called model S, which is simply a collection of isolated stations $j \in \mathcal{J}$, each consisting of N servers. The exogenous arrival flow at station j is Poisson, of rate $N\nu$, and the flows in different stations are independent. Upon arrival, a task chooses a queue in a given station at random and quits the system after completing service. The majorising property of model S is established in Theorem 1. The assertion of Theorem 1 has been independently proved by E. Thomas (unpublished), by using a modification of a method from [FMcD]; in fact, in the particular case where all probabilities p_j^* vanish, Theorem 1 gives the result of [FMcD].

Condition 1 reflects a popular point of view that, dealing with a load-balanced network, one has to assess the so-called dedicated traffic, through all sub-networks (including the network itself). The dedicated traffic is formed by the tasks that join a given collection of stations regardless of the state of the network, as opposite to the discretionary traffic where the decision to join a station depends on the state. Pictorially speaking, if the network is able to cope with the dedicated traffic, the discretionary traffic will be processed anyway. Note that in condition (1.1.1), $\zeta_{\mathcal{K}}$ can be considered as the rate of the dedicated traffic through set of stations $\mathcal{K}$ under an assumption that all servers in the network are busy (which is stressed in the form of the term $\sum\limits_{j\in\mathcal{J}} p_j^* p_{j,\mathfrak{m}}$).

Condition 1 indicates a domain in the space of parameters $\lambda_{\mathfrak{m}}$, p_i^* and $p_{i,\mathfrak{m}}$ which lies inside the *capacity domain*, i.e., the open set $\mathfrak{D}$ of the parameter values such that outside the closure $\bar{\mathfrak{D}}$ the network is super-critical, i.e., the Markov process is not recurrent. The problem of finding the capacity domain of an L-networks remains unresolved. A natural necessary condition (i.e., indicating a domain that is no smaller than $\mathfrak{D}$ is

Condition 2: there exist numbers $a^{(i)} \in [0,1)$, $i \in \mathcal{J}$, such that $\forall$ $\mathcal{K} \subset \mathcal{J}$,

$$\sum_{\mathfrak{m}:\, \mathfrak{m}\lceil\mathcal{K}^c=0} \left(\lambda_{\mathfrak{m}} + \sum_{j\in\mathcal{J}} a^{(j)} p_j^* p_{j,\mathfrak{m}}\right) \leq \sum_{i\in\mathcal{K}} a^{(i)}, \tag{1.3.1}$$

whereas for $\mathcal{K} = \mathcal{J}$,

$$\sum_{i\in\mathcal{J}} a^{(i)}(1-p_i^*) = \sum_{\mathfrak{m}} \lambda_{\mathfrak{m}}. \tag{1.3.2}$$

Eqn (1.3.2) simply means that the total arrival and departure rates are balanced. It is possible to check that condition 3 is sufficient for the positive recurrence under additional symmetry conditions (a nice example was found recently by F.I. Karpelevich (in preparation)).

There is a conjecture that for the case $J = 2$ (two stations) Condition 2 in fact suffices for positive recurrence of the Markov process in a general network L. For $N = 1$, $J = 2$, this conjecture was proved by I. Kurkova [Ku], following an approach developed in [FMM]. Geometrically, Condition 2 for $J = 2$ is that, inside the unit square $0 \leq a^{(1)}, a^{(2)} \leq 1$, the line $a^{(1)}(1-p_1^*) + a^{(2)}(1-p_2^*) = \sum_{\mathfrak{m}} \lambda_{\mathfrak{m}}$ has a non-empty intersection with two open half-planes $a^{(1)} > \sum\limits_{\mathfrak{m}:\, m_2=0} \left(\lambda_{\mathfrak{m}} + a^{(1)}p_{1,\mathfrak{m}} + a^{(2)}p_{2,\mathfrak{m}}\right)$ and $a^{(2)} > \sum\limits_{\mathfrak{m}:\, m_1=0} \Big(\lambda_{\mathfrak{m}} + a^{(1)}p_{1,\mathfrak{m}} + a^{(2)}p_{2,\mathfrak{m}}\Big)$.

Examples. To compare Conditions 2 and 3, we take $J = 2$. In the first example we assume that the inspection number distributions $\mathfrak{m}$ for which $\lambda_{\mathfrak{m}}$ or $p_{i,\mathfrak{m}}$ are non-zero, have $m_i \leq 1$. In other words, the network is determined by exogenous rates λ_1, λ_2 and $\lambda_{1,2}$ and probabilities p_i^* and $p_{i,1}$, $p_{i,2}$ and $p_{i,\{1,2\}}$, $i = 1, 2$. Here, $\lambda_i N$ is the exogenous dedicated arrival rate from outside in station i; the tasks arriving at this rate choose a queue in the

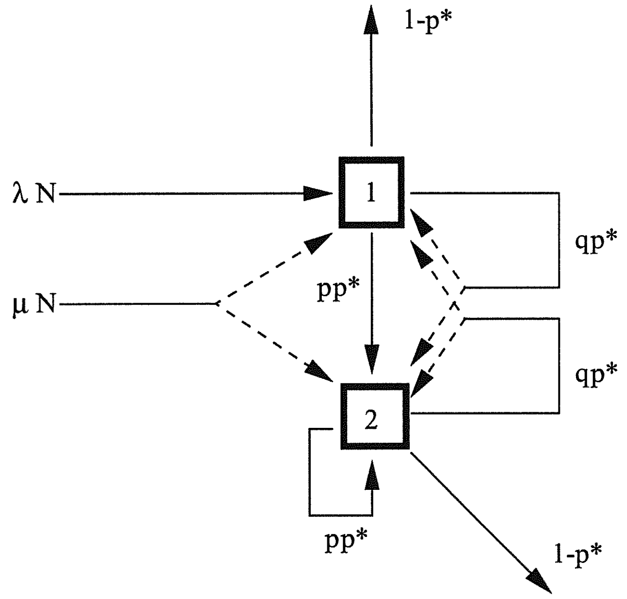

Figure 2

station at random. Next, $\lambda_{1,2}N$ is the discretionary arrival rate from outside (the tasks arriving at this rate choose at random one server in station 1 and one in station 2 and then join the shorter queue). Similarly, $p_{i,j}$ is the probability that after completing service in station i and deciding not to quit, the task will enter station j and join a randomly selected queue, while $p_{i,\{1,2\}}$ the probability that it will choose a server at random in each station and then join the shorter queue. To further simplify the matter, consider a particular case where $\lambda_2 = 0$ and set $\lambda_1 = \lambda$ and $\lambda_{1,2} = \mu$. Also assume that $p_1^* = p_2^* = p^*$, $p_{1,1} = p_{2,1} = 0$, $p_{1,2} = p_{2,2} = p$ and $p_{1,\{1,2\}} = p_{2,\{1,2\}} = q$, with $p + q = 1$. In this network, the exogenous arrivals go to station 1 or become discretionary while the re-entering tasks go to station 2 or become discretionary with probabilities that do not depend on the station where the previous service was completed. See Figure 2.

Conditions 1, 2 in this case take the following form:

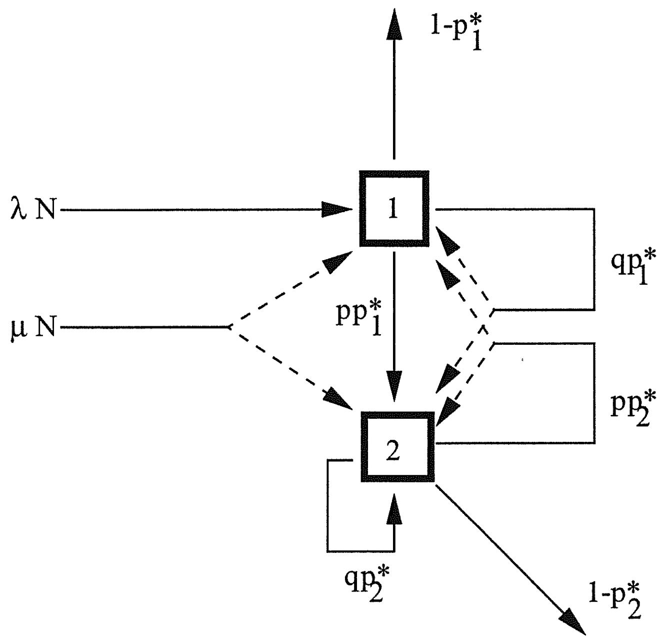

Figure 3

Condition 1:

$$\lambda < 1,\ \lambda + \mu < 2(1-p^*),\ pp^* < 1/2, \tag{1.4.1}$$

Condition 2:

$$\lambda < 1,\ \lambda + \mu < 2(1-p^*),\ \lambda + \mu < (p^*p)^{-1}(1-p^*). \tag{1.4.2}$$

If $p^*p < 1/2$, Conditions 1 and 2 coincide and hence Condition 1 describes the capacity domain of the network.

A modified version of this example is where p_1^* and p_2^* are not necessarily equal, while $p_{1,1} = p_{2,1} = 0$, $p_{1,2} = p_{2,\{1,2\}} = p$ and $p_{1,\{1,2\}} = p_{2,2} = q$, with $p + q = 1$. Here, as before the exogenous arrivals go to station 1 or become discretionary whereas the re-entering tasks go to station 2 or become discretionary with probabilities that depend on the last visited station, but in a symmetric way. See Figure 3. Here, in terms of parameters λ, μ, p_1^*, p_2^*, p and q, the domains given by Condition 1 and 2 are as follows.

Condition 1:

$$\lambda < 1,\ \lambda + \mu < 2 - p_1^* - p_2^*,\ p_1^* p + p_2^* q < 1, \tag{1.5.1}$$

Condition 2:

$$\lambda < 1,\ \lambda + \mu < (1 - p_1^*)\frac{1 - p_2^* p}{p_1^* q} + (1 - p_2^*). \tag{1.5.2}$$

As $p_1^* p + p_2^* q < 1$ implies $(p_1^* p)^{-1}(1 - p_2^* q) > 1$, Condition 1 is strictly more restrictive than 2.

2. The limit $N \to \infty$

Now we pass to limiting properties of the networks under consideration, as $N \to \infty$. The key fact is that the Markov process describing the L-network converges, as $N \to \infty$, to a deterministic process whose trajectory is given by a solution of a countable non-linear system of ordinary differential-difference equations. See Theorem 2. We write this system for an array of functions $\mathbf{u}(t) = \big(u_i(t;n)\big)$, where $t \geq 0$, $i \in \mathcal{J}$ and $n \in \mathbb{Z}_+$:

$$\dot{u}_i(t;n) = u_i(t;n+1) - u_i(t;n) + \sum_{\mathfrak{m}} \Big(\lambda_{\mathfrak{m}} + \sum_{j \in \mathcal{J}} u_j(t;1) p_j^* p_{j,\mathfrak{m}}\Big) \tag{2.1}$$

$$\times \sum_{\substack{\mathfrak{m}', \mathfrak{m}''\,:\, m_i' \geq 1, \\ \mathfrak{m}' + \mathfrak{m}'' = \mathfrak{m}}} \frac{m_i'}{|\mathfrak{m}'|} \prod_{l \in \mathcal{J}} \binom{m_l}{m_l'} \Big(u_l(t;n-1) - u_l(t;n)\Big)^{m_l'} \Big(u_l(t;n)\Big)^{m_l''}.$$

Here and below, the sum $\mathfrak{m}' + \mathfrak{m}''$ of inspection number distributions is understood component-wise. The summation $\sum\limits_{\mathfrak{m}',\mathfrak{m}'':\, m_i' \geq 1,\ \mathfrak{m}'+\mathfrak{m}''=\mathfrak{m}}$ includes the case $\mathfrak{m}' = \mathfrak{m}$, i.e., $\mathfrak{m}'' = 0$.

The meaning of Eqn (2.1) is that the change in $u_i(t;n)$, the (limiting) portion of the queues of length $\geq n$ in station i at time t may be produced when (i) a task completes service, which is described by the term $u_i(t;n+1) - u_i(t;n)$ and (ii) a task arrives (from outside or within the network) and joins a queue of length $n-1$. The latter situation is analysed by taking into account all possibilities, first at the level of rates $\lambda_{\mathfrak{m}}$ and $\sum_{j\in\mathcal{J}} u_j(t;1)p_j^* p_{j,\mathfrak{m}}$ and then by specifying the composition of the sample inspected. E.g., given $\mathfrak{m}'$ and $\mathfrak{m}''$, the sample contains m_l' queues of length $n-1$ and m_l'' of length $\geq n$ from station $l \in \mathcal{J}$; the condition $m_i' \geq 1$ is necessary here.

The following boundary condition is observed at $n=0$:

$$u_i(t;0) \equiv 1, \quad t \geq 0, \ i \in \mathcal{J}. \tag{2.2.1}$$

We also impose an initial condition $\mathbf{u}(0) = \mathbf{g}$, or, component-wise,

$$u_i(t;n)\big|_{t=0} = g_i(n), \ i \in \mathcal{J}. \ n \in \mathbb{Z}_+. \tag{2.2.2}$$

Here sequence $\big(g_i(n)\big)$ obeys $1 = g_i(0) \geq g_i(1) \geq \ldots \geq 0$ and, possibly, $\sum_{n\in\mathbb{Z}_+} g_i(n) < \infty$, $i \in \mathcal{J}$. Similar properties are expected from $\mathbf{u}(t)$, $t > 0$.

In Theorem 2 we establish a (global) existence and uniqueness of a solution $\mathbf{u}(t;\mathbf{g})$ to the initial-boundary value problem (2.1), (2.2.1-2). Here, and below, $\mathbf{g}$ is the array of initial data $(g_i(n),\ n \in \mathbb{Z}_+,\ i \in \mathcal{J})$ and $\mathbf{u}(t,\mathbf{g})$ is the array of functions $(u_i(t;n),\ n \in \mathbb{Z}_+,\ i \in \mathcal{J})$ satisfying (2.1), (2.2.1-2). Next, in Theorem 3 we prove the convergence of the r-process in model L to this solution as $N \to \infty$.

Next, we analyse a fixed point of (2.1), (2.2.1), i.e., an array $(a_i(n),\ i \in \mathcal{J}$, $n \in \mathbb{Z})$ obeying $1 = a_i(0) \geq a_i(1) \geq \ldots \geq 0$ and $\sum_{n\geq 1} a_i(n) < \infty$, $i \in \mathcal{J}$, and

$$a_i(n) - a_i(n+1) = \sum_{\mathfrak{m}} \Big(\lambda_{\mathfrak{m}} + \sum_{j\in\mathcal{J}} a_j(1)p_j^* p_{j,\mathfrak{m}}\Big)\times \tag{2.3}$$

$$\sum_{\substack{\mathfrak{m}',\mathfrak{m}'':\ m_i' \geq 1,\\ \mathfrak{m}'+\mathfrak{m}''=\mathfrak{m}}} \frac{m_i'}{|\mathfrak{m}'|}\prod_{l\in\mathcal{J}}\binom{m_l}{m_l'}\Big(a_l(n-1)-a_l(n)\Big)^{m_l'}\Big(a_l(n)\Big)^{m_l''}.$$

Observe that $a_j(n) \equiv 1$ is always a fixed point: it corresponds to a 'saturated' regime where all queue-lengths are set to be infinite. In Theorem 4 we prove that if there exists a fixed point $\mathbf{a}$ with $\sum_{j\in\mathcal{J},n\in\mathbb{Z}_+} a_j(n) < \infty$ (which is the case under condition (1.1)) then it has a 'global' attracting property. Finally, in Theorem 5 we establish the existence of such a fixed point $\mathbf{a}$.

An important role in the analysis is played by the quantity

$$V(t,\mathbf{g};n)=\sum_{1\leq i\leq J}\sum_{n'\geq n}u_i(t,\mathbf{g};n'),\ t\geq 0,\ n\in\mathbb{Z}_+. \tag{2.4}$$

Assuming that $\sum_{n\in\mathbb{Z}_+}g(n)<\infty$, a straightforward algebra leads to the following formula for the derivative $\dot{V}(t,\mathbf{g};n)$:

$$\dot{V}(t,\mathbf{g};n)=-\sum_{i\in\mathcal{J}}u_i(t,\mathbf{g};n)+\sum_{\mathrm{m}}\left(\lambda_{\mathrm{m}}+\sum_{j\in\mathcal{J}}u_j(t,\mathbf{g};1)p_j^*p_{j,\mathrm{m}}\right)$$

$$\times\prod_{l\in\mathcal{J}}\big(u_l(t,\mathbf{g};n-1)\big)^{m_l},\ t\geq 0,\ n\geq 1, \tag{2.5}$$

with the initial condition

$$V(0,\mathbf{g};n)=\sum_{i\in\mathcal{J}}\sum_{n'\geq n}g_i(n'),\ n\in\mathbb{Z}_+. \tag{2.6}$$

In particular, by using Lemma 2.1 we establish that

$$V(t,\mathbf{g};n)\leq Z(t,\mathbf{g};n),\ t\geq 0,\ n\in\mathbb{Z}_+, \tag{2.7}$$

where $Z(t,\mathbf{g};n)$ satisfies the linear system

$$\dot{Z}(t,\mathbf{g};n)=\nu\big(Z(t,\mathbf{g};n-1)-Z(t,\mathbf{g};n)\big)-Z(t,\mathbf{g};n)+Z(t,\mathbf{g};n+1), \tag{2.8}$$

with the initial and boundary conditions

$$Z(0,\mathbf{g};n)=\sum_{i\in\mathcal{J}}\sum_{n'\geq n}g_i(n'),\ n\in\mathbb{Z}_+,\ Z(t,\mathbf{g};0)-Z(t,\mathbf{g};1)=J,\ t\geq 0, \tag{2.9}$$

and

$$\lim_{n\to\infty}Z(t,\mathbf{g};n)=0,\ t\geq 0. \tag{2.10}$$

Here ν is the quantity defined in (1.2.1); when $\nu<1$ system (2.8), (2.9) (which corresponds to a collection of J isolated stations) has a unique fixed point $B=(B(n),\ n\in\mathbb{Z}_+)$, satisfying

$$B(n)-B(n+1)=\nu\big(B(n-1)-B(n)\big),\ n\in\mathbb{Z}_+. \tag{2.11}$$

This fixed point is of the form $B(n) = J\nu^n(1-\nu)^{-1}$ and has a global attracting property in a sense similar to above. Bound (2.7) helps to establish similar facts (existence and the global attraction) for the original system (2.1), (2.2.1-2); it is also used in the proof of Theorem 2.

Another key property is that a fixed point $\mathbf{a} \in \mathcal{U}^{\mathcal{J}}$ obeys

$$\sum_{i\in\mathcal{J}} a_i(n) = \sum_{\mathrm{m}} \Big(\lambda_{\mathrm{m}} + \sum_{j\in\mathcal{J}} a_j(1)p_j^* p_{j,\mathrm{m}}\Big) \prod_{l\in\mathcal{J}} \big(a_l(n-1)\big)^{m_l},\ n \geq 1. \tag{2.12}$$

In particular, for $n = 1$,

$$\sum_{i\in\mathcal{J}} a_i(1)(1-p_i^*) = \Lambda, \tag{2.13}$$

where

$$\Lambda = \sum_{\mathrm{m}} \lambda_{\mathrm{m}}. \tag{2.14}$$

We use Eqn (2.12) to study the decay of $a_j(n)$ for large n. Theorem 2 directly implies an inequality between $a_j(n)$, and $B(n)$ (see Eqn (3.4) in Theorem 6(A)) which provides an exponential bound for $a_j(n)$. However, a more interesting *super-exponential* bound can be proved. Assume the following

Condition 3:

$$\vartheta = \sup\Bigg[\max_{\mathcal{K}\subseteq\mathcal{J}} \frac{1}{\#\mathcal{K}} \sum_{\mathrm{m}:\mathrm{m}\lceil\mathcal{K}^c=0} \Big(\lambda_{\mathrm{m}} + \sum_{j\in\mathcal{J}} b_j p_j^* p_{j,\mathrm{m}}\Big) :$$

$$b_1, \dots, b_J \in [0,1],\ \sum_{1\leq j\leq J} b_j(1-p_j^*) = \Lambda\Bigg] < 1. \tag{2.15}$$

Then (see Theorem 5), for any $\vartheta_1 \in (\vartheta, 1)$ there exists a constant $C > 0$ (that can be assessed numerically) such that:

$$a_j(n) \leq C\vartheta_1^{(m_*^n-1)/(m_*-1)},\ i \in \mathcal{J},\ n \in \mathbb{Z}_+, \tag{2.16}$$

where

$$m_* = \min\Bigg[|\mathrm{m}| :\ \lambda_{\mathrm{m}} + \sum_{j\in\mathcal{J}} p_j^* p_{j,\mathrm{m}} > 0\Bigg]. \tag{2.17}$$

Condition 3 is weaker than 1 as the maximum in (2.15) is over a smaller set than in (1.1). Eqn (2.16) produces a super-exponential bound when $m_* \geq 2$, i.e., the task always inspects more than one queue before it joins one (the bound remains valid when $m_* = 1$ and $(m_*^n - 1)/(m_* - 1)$ is defined to be n). This is an analogue of the aforementioned results from [VDK] and [MS] on the super-exponential decay of the tail histograms, hence the term fast Jackson-type networks in the title of this paper.

3. Formal preliminaries and main theorems

The state space of the q-process for model L is a Cartesian product $\left(\mathbb{Z}_+^N\right)^{\mathcal{J}}$ and that of the r-process $\mathcal{U}_N^{\mathcal{J}}$. Here, $\mathcal{U}_N$ is the collection of the sequences $r = (r(n), n \in \mathbb{Z}_+)$ such that $1 = r(0) \geq r(1) \geq r(2) \geq \ldots \geq 0$, $r(n)N$ is integer and $r(n) = 0$ for n large enough. We denote by $Q_{j,r}(L,t)$ and $Q_{j,r}(S,t)$, $t \geq 0$, $j \in \mathcal{J}$, $1 \leq r \leq N$, the random variables forming the q-process in models L and S; $\mathbf{T}_N(t)$ denotes the transition semi-groups for the r-process in model L.

Theorem 1. *$\forall$ $N \geq 1$ and $x \geq 0$, if* $\sum\limits_{j\in\mathcal{J}} \sum\limits_{1\leq r\leq N} Q_{j,r}(L,0) = \sum\limits_{j\in\mathcal{J}} \sum\limits_{1\leq r\leq N} Q_{j,r}(S,0)$ *then* $\forall$ $t \geq 0$

$$\mathbb{E}\left(\sum_{j\in\mathcal{J}} \sum_{1\leq r\leq N} Q_{j,r}(L,t) - x\right)_+ \leq \mathbb{E}\left(\sum_{j\in\mathcal{J}} \sum_{1\leq r\leq N} Q_{j,r}(S,t) - x\right)_+ . \qquad (3.1)$$

Theorem 1 establishes a majorant in the sense $\leq_{\rm c}$ (see, e.g., [S], Sect. 1.3 and Chapters 5–7), popular in the queueing theory context. It implies that if, for a given N, model S has a stable equilibrium regime, so does L.

Corollary. *Under Condition* 1 (see (1.1)) *model L has a unique equilibrium distribution.*

Now turn to the limit $N \to \infty$. The state space of the limiting r-process is denoted by $\mathcal{U}^{\mathcal{J}}$: as before, it is the Cartesian product of J copies of $\mathcal{U}$, the space of the limiting tail histograms. A point of $\mathcal{U}$ is a sequence $r = \big(r(n)\big)$ where $1 = r(0) \geq r(1) \geq \ldots \geq 0$ and $\sum_n r(n) < \infty$. It is also convenient to consider a larger space $\bar{\mathcal{U}}^{\mathcal{J}}$ where $\bar{\mathcal{U}}$ consists of sequences $r = \big(\big(r(n)\big)$ where

$1 = r(0) \geq r(1) \geq \ldots \geq 0$. In probabilistic terms, $\mathcal{U}$ contains probability measures on $\mathbb{Z}_+$ with a finite expectation, while $\bar{\mathcal{U}}$ is formed by probability measures on the extended set $\mathbb{Z}_+ \cup \{\infty\}$. Points of $\mathcal{U}^{\mathcal{J}}$ and $\bar{\mathcal{U}}^{\mathcal{J}}$ are denoted, as before, by bold symbols, and referred to as arrays (e.g., $\mathbf{r} = \big(r_i(n),\ i \in \mathcal{J},\ n \in \mathbb{Z}_+\big)$).

The norm $||\mathbf{r}|| = \sup_{i\in\mathcal{J}} \sup_{n\in\mathbb{Z}_+} \dfrac{|r_i(n)|}{n+1}$ makes $\bar{\mathcal{U}}^{\mathcal{J}}$ a complete compact metric space. The corresponding topology is understood when we refer to continuity and convergence in $\bar{\mathcal{U}}^{\mathcal{J}}$ and $\mathcal{U}^{\mathcal{J}}$.

Theorem 2. *For* $\forall\ \mathbf{g} \in \bar{\mathcal{U}}^{\mathcal{J}}$, *problem (2.1), (2.2.1-2) has a unique solution* $\mathbf{u}(t) = \mathbf{u}(t, \mathbf{g}), t \geq 0$, *in* $\bar{\mathcal{U}}^{\mathcal{J}}$. *If* $\mathbf{g} \in \mathcal{U}^{\mathcal{J}}$, $\mathbf{u}(t, \mathbf{g})$ *belongs to* $\mathcal{U}^{\mathcal{J}}$ $\forall\ t \geq 0$. *Furthermore,*

$$V(t, \mathbf{g}; n) \leq Z(t, \mathbf{g}; n), \tag{3.2}$$

where $Z(t, \mathbf{g}; n)$ *is a solution to* (2.8), (2.9).

The convergence result for finite times is contained in Theorem 3:

Theorem 3. *For* $\forall$ *continuous function* $f \colon \bar{\mathcal{U}}^{\mathcal{J}} \to \mathbb{R}$ *and* $t \geq 0$,

$$\lim_{N\to\infty} \sup_{\mathbf{g}\in\mathcal{U}_N^{\mathcal{J}}} |\mathbf{T}_N(t) f(\mathbf{g}) - f(\mathbf{u}(t, \mathbf{g}))| = 0, \tag{3.3}$$

and the convergence is uniform in t within bounded intervals.

Now consider properties of the fixed points of problem (2.1), (2.2.1). We are interested in the 'unsaturated' fixed points $\mathbf{a} = (a_j(n))$ that lie in $\mathcal{U}^{\mathcal{J}}$. A remarkable fact is that if such a point exists, it is unique and attracts the whole of $\mathcal{U}^{\mathcal{J}}$.

Theorem 4. *There exists at most one* $\mathbf{a} = (a_j(n)) \in \mathcal{U}^{\mathcal{J}}$ *solving* (2.3), *and if such a point exists then* $\forall\ \mathbf{g} \in \mathcal{U}^{\mathcal{J}}$, $\lim\limits_{t\to\infty} \mathbf{u}(t, \mathbf{g}) = \mathbf{a}$.

Furthermore, there exists at most one $B = (B(n))$ *solving* (2.11), *and if such a point exists then* $\forall\ \mathbf{g} \in \mathcal{U}^{\mathcal{J}}$, $\lim\limits_{t\to\infty} Z(t, \mathbf{g}) = B$.

Theorem 5. *Under Condition* 3 (see (2.15)) *there exists a solution* $\mathbf{a} = (a_j(n)) \in \mathcal{U}^{\mathcal{J}}$ *to* (2.3). *Furthermore, this solution satisfies super-exponential inequality* (2.16) *whenever* $m_* \geq 2$.

Theorem 6. *Under Condition* 1 (see (1.1)):

(A) *There exists a fixed point* $\mathbf{a} \in \mathcal{U}^{\mathcal{J}}$ *of system* (2.1), (2.2.1) *and a solution* $B = (B(n))$ *of system* (2.11). *Furthermore,* $B(n) = J\nu^n(1-\nu)^{-1}$ *and*

$$\sum_{j\in\mathcal{J}} \sum_{n\in\mathbb{Z}_+} a_j(n) \leq J(1-\nu)^{-1}. \tag{3.4}$$

(B) *The equilibrium distribution* π_N *converges as* $N \to \infty$ *to the measure concentrated at fixed point* $\mathbf{a}$.

4. A comparison between networks L and S

Proof of Theorem 1. The nonzero rates of transition $\mathbf{g} \mapsto \mathbf{g}'$, $\mathbf{g}, \mathbf{g}' \in \mathcal{U}_N^{\mathcal{J}}$, of the Markov r-process in model L are as follows.

$$\mathbf{g} \mapsto \mathbf{g} + \frac{1}{N}\mathbf{e}_i(n), \quad \text{rate } NA_{(i,n)}(\mathbf{g}), \tag{4.1.1}$$

$$\mathbf{g} \mapsto \mathbf{g} - \frac{1}{N}\mathbf{e}_i(n), \quad \text{rate } NB^{(0)}_{(i,n)}(\mathbf{g}), \tag{4.1.2}$$

$$\mathbf{g} \mapsto \mathbf{g} + \frac{1}{N}(\mathbf{e}_{i'}(n') - \mathbf{e}_i(n)), \quad \text{rate } NB^{(1)}_{(i,n)}(\mathbf{g})C_{i,(i',n')}(\mathbf{g} - \frac{1}{N}\mathbf{e}_i(n)). \tag{4.1.3}$$

Here, $\mathbf{e}_i(n)$, $n \in \mathbb{Z}_+$, $i \in \mathcal{J}$, stands for an array whose only non-zero component is assigned to station i and the value of the argument n, and addition of arrays is component-wise. Physically, (4.1.1), corresponds to an exogenous arrival in and (4.1.2) to the departure from the network, while (4.1.3) corresponds to the transfer of a task from one queue to another. Furthermore, for $i, j \in \mathcal{J}$, $n \in \mathbb{Z}_+$,

$$A_{(i,n)}(\mathbf{g}) = \sum_{\mathfrak{m}} \lambda_{\mathfrak{m}} \sum_{\substack{\mathfrak{m}', \mathfrak{m}'' :\ m'_i \geq 1, \\ \mathfrak{m}' + \mathfrak{m}'' = \mathfrak{m}}} \frac{m'_i}{|\mathfrak{m}'|}$$

$$\times \prod_{l\in\mathcal{J}} \binom{m_l}{m'_l} \Big(g_l(n-1) - g_l(n)\Big)^{m'_l} \Big(g_l(n)\Big)^{m''_l}, \tag{4.2.1}$$

$$B^{(0)}_{(i,n)}(\mathbf{g}) = \Big(g_i(n) - g_i(n+1)\Big)(1 - p^*_i), \tag{4.2.2}$$

$$B^{(1)}_{(i,n)}(\mathbf{g}) = \Big(g_i(n) - g_i(n+1)\Big)p^*_i, \tag{4.2.3}$$

$$C_{i,(j,n)}(\mathbf{g}) = \sum_{\mathfrak{m}} p_{i,\mathfrak{m}} \sum_{\substack{\mathfrak{m}', \mathfrak{m}'' :\, m'_j \geq 1,\\ \mathfrak{m}' + \mathfrak{m}'' = \mathfrak{m}}} \frac{m'_j}{|\mathfrak{m}'|}$$

$$\times \prod_{l \in \mathcal{J}} \binom{m_l}{m'_l} \Big(g_l(n-1) - g_l(n) \Big)^{m'_l} \Big(g_l(n) \Big)^{m''_l}. \qquad (4.2.4)$$

We want to examine the action of the generator $\mathbf{A}_N$ of the Markov r-process in network L on the functions $\mathbf{g} \in \mathcal{U}_N^{\mathcal{J}} \mapsto V(0, \mathbf{g}; n)$, see (2.6) (the argument 0 in this notation will be omitted).

Lemma 4.1. *The following formula holds:*

$$\mathbf{A}_N V(\mathbf{g}; n) = \sum_{\mathfrak{m}} \Big(\lambda_{\mathfrak{m}} + \sum_{j \in \mathcal{J}} g_j(1) p^*_j p_{j,\mathfrak{m}} \Big) \prod_{l \in \mathcal{J}} \big(g_l(n-1) \big)^{m_l} - \sum_{i \in \mathcal{J}} g_i(n). \quad (4.3)$$

Proof of Lemma 4.1. The lemma claims that $\mathbf{A}_N V(\mathbf{g}; n) = V^{(0)}(\mathbf{g}; n) + V^{(1)}(\mathbf{g}; n) + V^{(2)}(\mathbf{g}; n)$, where

$$V^{(0)}(\mathbf{g}; n) = \sum_{\mathfrak{m}} \lambda_{\mathfrak{m}} \prod_{l \in \mathcal{J}} \big(g_l(n-1) \big)^{m_l},$$

$$V^{(1)}(\mathbf{g}; n) = - \sum_{i \in \mathcal{J}} g_i(1),$$

$$V^{(2)}(\mathbf{g}; n) = \sum_{\mathfrak{m}} \sum_{j \in \mathcal{J}} g_j(1) p^*_j p_{j,\mathfrak{m}} \prod_{l \in \mathcal{J}} \big(g_l(n-1) \big)^{m_l}.$$

In fact, under the action of $\mathbf{A}_N$, the value of function $V(\mathbf{g}; n)$ can only increase or decrease by $1/N$ which corresponds to an arrival or departure of a task from a queue of length $\geq n$. From this point of view, the term $V^{(0)}(\mathbf{g}; n)$ describes the effect of an exogenous arrival, $V^{(1)}(\mathbf{g}; n)$ that of a potential departure (when service is completed) and $V^{(1)}(\mathbf{g}; n)$ that of a return. This completes the proof.

Our next step is Lemma 4.2 below:

Lemma 4.2. $\forall\ r_1, \ldots, r_J \in [0, 1]$,

$$\sum_{\mathfrak{m}} \Big(\lambda_{\mathfrak{m}} + \sum_{j \in \mathcal{J}} g_j(1) p^*_j p_{j,\mathfrak{m}} \Big) \prod_{l \in \mathcal{J}} r_l^{m_l}$$

$$\leq \max_{\mathcal{K}\subseteq\mathcal{J}} \sum_{\mathrm{m}:\,\mathrm{m}\lceil\mathcal{K}^c=0} \left(\lambda_{\mathrm{m}} + \sum_{j\in\mathcal{J}} g_j(1)p_j^*p_{j,\mathrm{m}}\right) \sum_{k\in\mathcal{J}} r_k. \tag{4.4}$$

Proof of Lemma 4.2. In view of the convexity of the LHS of (4.4), it is enough to check that (4.4) holds on the boundary of $[0,1]^{\mathcal{J}}$, where one of the r_j's takes value 0 or 1. Here, the problem is reduced to $J-1$ variables. Applying the same argument, it suffices to check (4.4) on the lower-dimensional parts of the boundary, etc. Finally, our task is reduced to checking (4.4) when some of the r_j's are 0's and the rest are 1's. Here, it is straightforward.

Back to the proof of Theorem 1, Lemmas 4.1 and 4.2 allow us to write the inequality

$$\mathbf{A}_N V(\mathbf{g};n) \leq \nu \sum_{i\in\mathcal{J}} g_i(n-1) - \sum_{i\in\mathcal{J}} g_i(n). \tag{4.5}$$

Observe that we replaced the factors $g_j(1)$ in the sum $\sum_{j\in\mathcal{J}} g_j(1)p_j^*p_{j,\mathrm{m}}$ in the RHS of (4.3) by 1. We use the notation $V_N(t,\mathbf{g};n)$ for the function $\Big(\mathbf{T}_N(t)V(\,\cdot\,;n)\Big)(\mathbf{g})$, the result of the action of the transition operator $\mathbf{T}_N(t)$ $=$ exp $\big(t\mathbf{A}_N\big)$ on $V(\mathbf{g};n)$. In other words,

$$NV_N(t,\mathbf{g};n) = \mathbb{E}\left(\sum_{j\in\mathcal{J}}\sum_{1\leq r\leq N} Q_{j,r}(L,t) - n\right)_+ .$$

Then

$$\begin{aligned}\dot{V}_N(t,\mathbf{g};n) &= \Big(\mathbf{T}_N(t)\mathbf{A}_N V(\,\cdot\,;n)\Big)(\mathbf{g})\\ &\leq \nu\Big(\mathbf{T}_N(t)Y(\,\cdot\,;n-1)\Big)(\mathbf{g}) - \Big(\mathbf{T}_N(t)Y(\,\cdot\,;n)\Big)(\mathbf{g}),\end{aligned} \tag{4.6}$$

where ν is defined in (1.2.1) and

$$Y(\mathbf{g};n) = \sum_{j\in\mathcal{J}} g_j(n). \tag{4.7}$$

Bound (4.6) shows that $V_N(t,\mathbf{g};n) \leq Z(t;\mathbf{g};n)$ where $Z(t;\mathbf{g};n)$ is the solution to problem (2.8), (2.9). But (2.8), (2.9) is just the system of equations for the expected values in model S:

$$NZ(t,\mathbf{g};n) = \mathbb{E}\left(\sum_{j\in\mathcal{J}}\sum_{1\leq r\leq N} Q_{j,r}(S,t) - n\right)_+ .$$

The proof of Theorem 1 is now complete.

5. Analysis of the limiting model

Proof of Theorem 2. The proof of the statements when the initial date $\mathbf{g} \in \mathcal{U}^{\mathcal{J}}$ is rather standard and may be done as in [VDK] or [MS]; both methods use a kind of monotonicity argument. We therefore omit the bulk of technical details. However, we note the following monotonicity property of the solution $\mathbf{u}(t)$:

Lemma 5.1. *If* $\mathbf{g} \geq \mathbf{g}'$ *then,* $\forall\, t \geq 0$,

$$\mathbf{u}(t, \mathbf{g}) \geq \mathbf{u}(t, \mathbf{g}') \tag{5.1}$$

(the inequalities between arrays are understood component-wise).

Proof of Lemma 5.1. It suffices to check that the RHS of (2.1) is monotone in all variables $u_k(t, \widetilde{n})$, $k \in \mathcal{J}$, $\widetilde{n} = n-1, n$. For $\widetilde{n} = n-1$ this is plain; for $\widetilde{n} = n$ it follows from a straightforward calculation.

A useful observation providing the proof of a part of Theorem 2 is related to the case where $\mathbf{g} \in \mathcal{U}^{\mathcal{J}}$ is as follows. According to (2.5), the derivative $\dot{V}(t, \mathbf{g}; n)$ is bounded as $\mathbf{u}(t, \mathbf{g})$ belongs to $\bar{\mathcal{U}}^{J}$. Thus, $V(t, \mathbf{g}; n)$ grows at most linearly with time. Therefore, if $\mathbf{g} \in \mathcal{U}^{\mathcal{J}}$, $\mathbf{u}(t, \mathbf{g})$ belongs to $\mathcal{U}^{J}$ $\forall\, t \geq 0$.

Inequality (3.2) is just a limiting form of (3.1) and follows from Theorem 3. This completes the proof of Theorem 2.

The proof of Theorem 3 is carried in Section 6.

Proof of Theorem 4. Consider the initial condition $\mathbf{g}^0$ with $g_i^0(n) = \delta_{0,n}$, $i \in \mathcal{J}$. Due to monotonicity property (5.1), the solution $\mathbf{u}(t, \mathbf{g}^0)$ is monotone non-decreasing with t and hence has a limit (in $\bar{\mathcal{U}}^{\mathcal{J}}$). Denote this limit by $\mathbf{a}^0$; then by continuity, $\mathbf{a}^0$ satisfies (2.3). Our aim is to show that if $\mathbf{a} \in \mathcal{U}^{\mathcal{J}}$ then $\mathbf{a}^0$ attracts any solution $\mathbf{u}(t, \mathbf{g})$ with $\mathbf{g} \in \mathcal{U}^{\mathcal{J}}$; this will imply the uniqueness of the fixed point in $\mathcal{U}^{\mathcal{J}}$.

Observe that, owing to (5.1), if $\mathbf{g} \geq \mathbf{a}^0$ then $\mathbf{u}(t, \mathbf{g}) \geq \mathbf{a}^0$ $\forall\, t \geq 0$, and if $\mathbf{g} \leq \mathbf{a}^0$ then $\mathbf{u}(t, \mathbf{g}) \leq \mathbf{a}^0$ $\forall\, t \geq 0$. Now set $W(t; n) = W(t; \mathbf{g}, n) = \sum\limits_{i \in \mathcal{J}} \sum\limits_{n' \geq n} \left| u_i(t; n') - a_i^0(n') \right|$, $t \geq 0$, $n \in \mathbb{Z}_+$. Assuming that

$W(t;1)\Big|_{t=0} = \sum_{j\in\mathcal{J}}\sum_{n'\geq 1}\left|g_j(n')-a_j^0(n')\right| < \infty$, we will show that, $\forall\ n \in \mathbb{Z}_+$, $\int_0^\infty \mathrm{d}t\sum_{j\in\mathcal{J}}\left|u_j(t;n)-a_j^0(n)\right| < \infty$ implying that $\lim_{t\to\infty}\left|u_j(t;n)-a_j^0(n)\right| = 0$.

First, let us prove the assertion under an additional assumption that $\mathbf{g}\geq\mathbf{a}^0$. We can then omit the absolute value sign in the definition of $W(t;n)$. We begin with the remark that $W(t;1)$ stays bounded in t. In fact, with the help of (2.5), $\dot{W}(t;1) = -\sum_{i\in\mathcal{J}}(1-p_i^*)\big(u_i(t;1)-a_i^0(1)\big)$ which is ≤ 0. In other words, $W(t;1)$ is non-increasing in time. Now we are going to use induction in n: for $n=0$, the above assertion holds automatically, owing to the boundary condition. Assuming the induction hypothesis for $\leq n-1$, write, in view of (2.5):

$$\dot{W}(s;n) = -\sum_{i\in\mathcal{J}}\big(u_i(s;n)-a_i^0(n)\big) + \sum_{\mathrm{m}}\left[\left(\lambda_{\mathrm{m}}+\sum_{j\in\mathcal{J}}u_j(t;1)p_j^*p_{j,\mathrm{m}}\right)\right. \tag{5.2}$$

$$\left.\times\prod_{l\in\mathcal{J}}\big(u_l(t;n-1)\big)^{m_l} - \left(\lambda_{\mathrm{m}}+\sum_{j\in\mathcal{J}}a_j^0(1)p_j^*p_{j,\mathrm{m}}\right)\prod_{l\in\mathcal{J}}\big(a_l^0(n-1)\big)^{m_l}\right]$$

or, integrating,

$$W(t;n) = W(0;n) + \int_0^t(\text{the RHS of (5.2)})\,\mathrm{d}s. \tag{5.3}$$

The LHS of (5.3) is $\leq W(t;1)$ and hence is bounded in t. In the RHS of (5.3), the sum

$$\sum_{\mathrm{m}}\int_0^t\left[\left(\lambda_{\mathrm{m}}+\sum_{j\in\mathcal{J}}u_j(s;1)p_j^*p_{j,\mathrm{m}}\right)\prod_{l\in\mathcal{J}}\big(u_l(s;n-1)\big)^{m_l}\right.$$

$$\left.-\left(\lambda_{\mathrm{m}}+\sum_{j\in\mathcal{J}}a_j^0(1)p_j^*p_{j,\mathrm{m}}\right)\prod_{l\in\mathcal{J}}\big(a_l^0(n-1)\big)^{m_l}\right]\mathrm{d}s$$

converges as $t\to\infty$, owing to the induction hypothesis. Therefore, the remaining integral

$$\int_0^t\left[-\sum_{i\in\mathcal{J}}\big(u_i(s;n)-a_i^0(n)\big)\right]\mathrm{d}s$$

also converges as $t \to \infty$. This verifies the induction step.

The case where $\mathbf{g} \leq \mathbf{a}^0$ is analyzed in a similar fashion. In a general case we pass to $\mathbf{g}^+ = \max\ [\mathbf{g}, \mathbf{a}^0]$ and $\mathbf{g}^- = \min\ [\mathbf{g}, \mathbf{a}^0]$ (both operations max and min are understood component-wise) and use again the monotonicity of $\mathbf{u}(t, \mathbf{g})$ in $\mathbf{g}$. This completes the proof of Theorem 4 for model L.

Finally, the analysis of the fixed point B of linear system (2.8), (2.9), is performed in a standard way, and we do not dwell on it.

Proof of Theorem 5. As in Section 4, set $Y(n)$ $(= Y(\mathbf{a}; n)) = \sum_{j \in \mathcal{J}} a_j(n)$. We then have $Y(0) = J$. As to $Y(n)$, $n \geq 1$, in view of (2.3), (2.5) we have that

$$Y(n) - Y(n+1) = \sum_{\mathbf{m}} \left(\lambda_{\mathbf{m}} + \sum_{j \in \mathcal{J}} a_j(1) p_j^* p_{j,\mathbf{m}} \right)$$

$$\times \left[\prod_{l \in \mathcal{J}} \Big(a_l(n-1) \Big)^{m_l} - \prod_{l \in \mathcal{J}} \Big(a_l(n) \Big)^{m_l} \right].$$

This suggests that $Y(n)$ can be sought in the form

$$Y(n) = \sum_{\mathbf{m}} \left(\lambda_{\mathbf{m}} + \sum_{j \in \mathcal{J}} a_j(1) p_j^* p_{j,\mathbf{m}} \right) \prod_{l \in \mathcal{J}} \Big(a_l(n-1) \Big)^{m_l}, \qquad (5.4)$$

which, according to Lemma 4.2, is $\leq \vartheta Y(n-1)$, ϑ being given by (2.15). So, $Y(n) \leq J\vartheta^n$, and if $\vartheta < 1$, this implies an exponential and in fact a super-exponential decay of $a_i(n)$ as $n \to \infty$, i.e., bound (2.16). The proof of Theorem 5 is now complete.

Proof of Theorem 6(A). The existence, under Condition 1, of the fixed point $B = (B(n))$ is straightforward. Also, bound (3.1) implies that in model L, under (1.1), $\mathbf{a}^0 \in \mathcal{U}^{\mathcal{J}}$ and Eqn (3.4) holds. This completes the proof of statement (A).

The proof of Theorem 6(B) is carried in Section 6.

6. Convergence to the limiting model

Proof of Theorem 3. The proof of this theorem essentially repeats the argument used in [MS] (and other papers quoted in Introduction), and we will only sketch it. We use notation similar to [MS], Section 3. The first step is to make a statement about the convergence of the generator $\mathbf{A}_N$ of the Markov r-process in model L (see Section 4) to the operator related to the RHS of of (2.1) precise. The following lemma is used here:

Lemma 6.1. *$\forall$ $\mathbf{g} \in \bar{\mathcal{U}}^{\mathcal{J}}$, $t \geq 0$, $j \in \mathcal{J}$ and $N \in \mathbb{Z}$, the derivatives $\dfrac{\partial \mathbf{u}(t,\mathbf{g})}{\partial g_j(n)}$, and $\dfrac{\partial^2 \mathbf{u}(t,\mathbf{g})}{\partial g_j(n) \partial g_{j'}(n')}$ exist and satisfy*

$$\left| \frac{\partial u_k(r,t,\mathbf{g})}{\partial g_j(n)} \right| \leq C_1 e^{C_2 t} \tag{6.1.1}$$

and

$$\left| \frac{\partial^2 u_k(r,t,\mathbf{g})}{\partial g_j(n) \partial g_{j'}(n')} \right| \leq C_1 e^{C_3 t}, \tag{6.1.2}$$

where C_1, C_2 and C_3 are positive constants.

The proof of Lemma 6.1 is similar to that of Lemma 3.2 from [MS] and omitted. We then introduce the set D of functions $f : \bar{\mathcal{U}}^{\mathcal{J}} \to \mathbb{R}$ for which the derivatives $\dfrac{\partial f(\mathbf{g})}{\partial g_j(n)}$, and $\dfrac{\partial^2 f(\mathbf{g})}{\partial g_j(n) \partial g_{j'}(n')}$ exist for all $\mathbf{g}, j, j', n, n'$ and are uniformly bounded in the absolute value by a constant $c = c(f) < \infty$. D is dense in the space $\mathrm{C}(\bar{\mathcal{U}}^{\mathcal{J}})$ of continuous functions on $\bar{\mathcal{U}}^{\mathcal{J}}$ (with the standard sup-norm). Furthermore, let $\mathbf{A}_N$ denote the generator of the r-process in model L, with the matrix entries given by (4.1). Then, as it is easy to see, $\forall$ $f \in D$, $\lim\limits_{N \to \infty} \mathbf{A}_N f(\mathbf{g}) = \mathbf{A} f(\mathbf{g})$, where $\mathbf{A}$ is an operator defined by

$$\mathbf{A} f(\mathbf{r}) = \sum_{1 \leq i \leq J} \sum_{n \geq 1} \Bigg[[r_i(n+1) - r_i(n)] + \sum_{\mathfrak{m}} \Bigg(\lambda_{\mathfrak{m}} + \sum_{1 \leq j \leq J} r_j(1) p_j^* p_{j,\mathfrak{m}} \Bigg) \times$$

$$\sum_{\substack{\mathfrak{m}', \mathfrak{m}'' :\, m_i' \geq 1, \\ \mathfrak{m}' + \mathfrak{m}'' = \mathfrak{m}}} \frac{m_i'}{|\mathfrak{m}'|} \prod_{l \in \mathcal{J}} \binom{m_l}{m_l'} \Big(r_l(n-1) - r_l(n) \Big)^{m_l'} \Big(r_l(n) \Big)^{m_l''} \Bigg] \frac{\partial}{\partial r_i(n)} f(\mathbf{r}). \tag{6.2}$$

Observe that the operator semi-groups $\mathbf{T}_N(t)$ and $\mathbf{T}(t)$, $t \geq 0$, generated in $\mathrm{C}(\bar{\mathcal{U}}^{\mathcal{J}})$ by $\mathbf{A}_N$ and $\mathbf{A}$ are continuous and contracting.

If D_0 denotes the subset of D consisting of functions f that depend on finitely many variables $g_j(n)$ then D_0 is dense in D and hence in $\mathrm{C}(\bar{\mathcal{U}}^{\mathcal{J}})$. As in [VDK] and [MS], it is easy to see that $\mathbf{T}(t)D_0 \subset D$. It remains to use general facts about the convergence of distributions of Markov processes (see [EK], Chapter 1, Proposition 3.3 and Theorem 7.1). This gives the assertion of Theorem 3.

Proof of Theorems 6(B). Theorem 1 allows us to use the same argument as, e.g., in [MS]. Namely, the sequence of probability measures P_N is compact, and any of its limit points is a measure concentrated on fixed points of (2.1), (2.2.1-2). Thus, it suffices to check that if π is a limit point then $\pi\left(\mathcal{U}^{\mathcal{J}}\right) = 1$, which in turn will follow from the bound $\mathbb{E}_\pi V(\,\cdot\,;1) < \infty$. Now by Theorem 1,

$$\mathbb{E}_\pi V(\,\cdot\,;1) \leq \sum_{j\in\mathcal{J}} \sum_{n\in\mathbb{Z}_+} \nu_j^n,$$

and the RHS is finite under condition (1.1). This completes the proof of Theorem 6(B).

7. A simplified model

The model considered in this section is where probabilities p_j^* and $p_{j,\mathfrak{m}}$ do not depend on j; thus subscript j will be omitted from this notation. The main simplification is that the total throughput rate in the whole network is $\Lambda(1-p^*)^{-1}$, where Λ is the sum (2.14). It is also easy to calculate the total rate $\Lambda_{\mathcal{K}}$ of the dedicated traffic in a sub-set of stations $\mathcal{K} \subseteq \mathcal{J}$ (cf. (1.2)):

$$\Lambda_{\mathcal{K}} = \sum_{\mathfrak{m}:\, m_j = 0 \forall\, j \notin \mathcal{K}} \left(\lambda_{\mathfrak{m}} + \Lambda p^*(1-p^*)^{-1} p_{\mathfrak{m}}\right). \qquad (7.1)$$

Thus, the above principle of the dedicated traffic capacity (see Section 1) can be now stated as a formal **Conjecture**: *the inequality*

$$\eta := \max_{\mathcal{K}} \frac{1}{\#\mathcal{K}} \Lambda_{\mathcal{K}} < 1 \qquad (7.2)$$

describes the sub-criticality domain for the simplified model L, in the sense that a) condition (7.2) is sufficient, and b) if the inequality sign in (7.2) is reversed, the network does not have a proper equilibrium distribution.

As was said, the general results given in the previous sections are not sufficient for proving this conjecture. However, under additional assumptions about rates λ_{m} and probabilities p_{m} we can establish this conjecture in the limit $N \to \infty$. These assumptions are that, for some $M^0, M^1 \geq 1$ and q_j^0, q_j^1, $\in [0,1]$, $j \in \mathcal{J}$, with $\sum_{j\in\mathcal{J}} q_j^0 = \sum_{j\in\mathcal{J}} q_j^1 = 1$,

$$\lambda_{\mathrm{m}} = \Lambda \prod_{l\in\mathcal{J}} (q_l^0)^{m_l} \binom{M^0}{m_1,\dots,m_J}, \text{ if } |\mathrm{m}| = M^0,$$

$$\lambda_{\mathrm{m}} = 0, \text{ otherwise}, \tag{7.3}$$

and

$$p_{\mathrm{m}} = \Lambda \prod_{l\in\mathcal{J}} (q_l^1)^{m_l} \binom{M^1}{m_1,\dots,m_J}, \text{ if } |\mathrm{m}| = M^1,$$

$$p_{\mathrm{m}} = 0, \text{ otherwise}. \tag{7.4}$$

Here, $\binom{M}{m_1,\dots,m_J}$, for $M, m_1, \dots, m_J \in \mathbb{Z}_+$, $\sum_{1\leq k\leq J} m_k = M$, stands for the product $\binom{M}{m_1}\binom{M-m_1}{m_2}\dots$. The L-model of this form is called multinomial. Here,

$$\Lambda_{\mathcal{K}} = \Lambda\left(\left(\sum_{l\in\mathcal{K}} q_l^0\right)^{M^0} + \frac{p^*}{1-p^*}\left(\sum_{l\in\mathcal{K}} q_l^1\right)^{M^1}\right). \tag{7.5}$$

Also, for a multinomial L-model, Eqn. (2.5) takes the form

$$\dot{V}(t,\mathbf{g};n) = -\sum_{1\leq i\leq J} u_i(t,\mathbf{g};n) + \Lambda\left(\sum_{1\leq k\leq J} q_k^0 u_k(t,\mathbf{g};n-1)\right)^{M^0}$$

$$+p^* \sum_{1\leq j\leq J} u_j(t,\mathbf{g};1)\left(\sum_{1\leq k\leq J} q_k^1 u_k(t,\mathbf{g};n-1)\right)^{M^1}, \tag{7.6}$$

and Eqn (2.12)

$$Y(n) = \Lambda\left(\sum_{1\leq k\leq J} q_k^0 a_k(n-1)\right)^{M^0} + \frac{p^*}{1-p^*}\Lambda\left(\sum_{1\leq k\leq J} q_k^1 a_k(n-1)\right)^{M^1}. \tag{7.7}$$

Here, as in Section 6, $Y(n) = \sum_{j\in\mathcal{J}} a_j(n)$, $n \in \mathbb{Z}_+$.

A condition for a super-exponential decay in the multinomial L-model is that

$$M_* = \min[M^0, M^1] \geq 2, \tag{7.8}$$

and

$$\eta_1 := \Lambda \left(\max_{k\in\mathcal{J}} \left(q_k^0\right)^{M^0} + \frac{p^*}{1-p^*} \max_{k\in\mathcal{J}} \left(q_k^1\right)^{M^1} \right) < 1. \tag{7.9}$$

Theorem 7. *For a multinomial L-model, under condition* (7.2) *there exists a solution* $\mathbf{a} \in \mathcal{U}^{\mathcal{J}}$ *of* (2.3). *On the contrary, if the inequality sign in* (7.2) *is reversed, system* (2.1) *does not have a fixed point in* $\mathcal{U}^{\mathcal{J}}$.

Furthermore, if conditions (7.2), (7.8) *and* (7.9) *holds, fixed point* $\mathbf{a}$ *in the multinomial model L has a super-exponential decay:*

$$a_i(n) \leq C\eta_1^{(M_*^n-1)/(M_*-1)}, \; n \in \mathbb{Z}_+, \; i \in \mathcal{J}. \tag{7.10}$$

Applying Theorem 4 yields the following

Corollary. *For a multinomial L-model, under condition* (7.2) *there exists a unique* $\mathbf{a} \in \mathcal{U}^{\mathcal{J}}$ *solving* (2.3), *and* $\forall$ $\mathbf{g} \in \mathcal{U}^{\mathcal{J}}$, *the solution* $\mathbf{u}(t, \mathbf{g})$ *of* (2.1), (2.2.1-2) *converges to* $\mathbf{a}$.

Proof of Theorem 7. As in the proof of Theorem 6(A), we analyse Eqn (7.7). A bound similar to (4.4) is that

$$\text{the RHS of (7.7)} \leq \eta Y(n-1) - Y(n), \; n \in \mathbb{Z}_+. \tag{7.11}$$

So, if $\eta < 1$, $Y(n)$ decays exponentially with n. However, if one in addition assumes (7.8) and (7.9), it is possible to obtain more:

$$Y(n) \leq \eta_1 Y(n-1)^{M_*}, \; n \in \mathbb{Z}_+.$$

which leads to (7.10).

Reversing the inequality sign in (7.2) leads to the absence of a fixed point in $\mathcal{U}^{\mathcal{J}}$ in a straightforward way.

In memoriam. We dedicate this paper to the memory of Roland Dobrushin (1929–1995) whose influence upon both of us, both scientifically and personally, is difficult to overestimate. Roland pioneered rigorous studying of many aspects of the queueing network theory; the interested reader can find more details in [KPS]. See also [D].

Acknowledgements. This work was partly supported by the Department of Statistics and Applied Probability, University of California, Santa Barbara, CA; the London Mathematical Society; St John's College, Cambridge; Dublin Institute for Advanced Studies; I.H.E.S., Bures-sur-Yvette; INRIA Rocquencourt, project 'Hypercom'; The French-Russian Lyapunov Institute, Project 6-97; Russian RFFI Grant 96-01-00150; EC Grant 'Training Mobility and Research' (Contract ERBMRXT-CT 960075A), and INTAS Grant 'Mathematical Methods for Stochastic Discrete Event Systems' (INTAS 93-820). The authors would like to thank V. Anantharam, Y. Baryshnikov, O. Boxma, P. Dellar, M. Harrison, P. Jacquet, F. Karpelevich, F. Kelly, I. Kurkova, J. Lewis, J. Martin, V. Priezzhev, A. Rybko, A. Stolyar, D. Stuart, E. Thomas, V. Vol'pert, J. Wessels and R. Williams for valuable comments.

References

[AWZ] I.J. Aidan, J. Wessels and W.H.M. Ziim. Analysis of the asymmetric shortest queue problem. *Queueing Sysytems*, **8** (1996), 1–58.

[D] R.L Dobrushin. Switching networks, Gibbsian fields – interconnections. In: *Proc. 1st World Congress of the Bernoulli Society.* Tashkent, 1986. Utrecht: VN Sci. Press, 1987, pp. 377–393.

[EK] S. Ethier and T. Kurtz. *Markov Processes: Characterization and Convergence.* New York, Wiley, 1986.

[FIM] G. Fayolle, R. Iasnogorodski and V. Malyshev. *Random Walks in the Quarter-Plane.* Berlin: Springer-Verlag, 1999.

[FMM] G. Fayolle, V.A. Malyshev and M.V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains.* Cambridge: CUP, 1995.

[FMcD] R. Foley, D. McDonald. Join the shortest queue: stability and exact asymptotics. Preprint, Georgia Institute of Technology and University of Ottawa, 1999.

[FC] S.Foss and N.Chernova. On the stability of a partially accessible queue with state-dependent routing. *Queueing Systems*, **29** (1998), 55–73.

[KPS] F.I. Karpelevich, E.A. Pechersky and Yu.M. Suhov. Dobrushin's approach to queueing network theory, *J. Appl. Math. Stoch. Anal.*, **9** (1996), 373–398.

[Ke] F.P. Kelly. Loss networks. *Ann. Appl. Prob.*, **1** (1991), 319–378.

[Ku] I. Kurkova. A load-balanced network with two servers. Technical Report, EURANDOM, University of Eindhoven, 1999.

[MS] J.B. Martin and Yu.M. Suhov. Fast Jackson networks, To appear in *Ann. Appl. Prob.*, 1999.

[M] M. Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*, PhD Thesis, University of California, Berkeley, 1996.

[MV] M.Mitzenmacher and B.Voecking. The asymptotics of selecting the shortest of two, improved. Technical Report, Harvard University, 1999.

[S] D. Stoyan. *Comparison Methods for Queues and Other Stochastic Models*. New York et al.: John Wiley and Sons, 1983.

[T] S.R.E. Turner. The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences*, **12** (1998), 109–124.

[VDK] N.D. Vvedenskaya, R.L Dobrushin and F.I. Karpelevich. A queueing system with selection of the shortest of two queues: an asymptotical approach, *Problems of Information Transmission*, **32** (1996), 15–27.

[VS] N.D. Vvedenskaya and Yu.M. Suhov. Dobrushin's mean-field approximation for a queue with dynamic routeing, *Markov Proc. Rel. Fields*, **3** (1997), 493–527.