FAST JACKSON NETWORKS

J. B. MARTIN[1] and Yu. M. SUHOV[2]

University of Cambridge and Ecole Normale Supérieure, Paris, and University of Cambridge and IPIT, Russian Academy of Sciences

We extend the results of [8] to Jackson networks. Each node j, $1 \le j \le J$ of the network consists of N identical channels, each with an infinite buffer and a single server with service rate μ_j. The network is fed by a family of independent Poisson flows of rates $N\lambda_1, \dots, N\lambda_J$ arriving at the corresponding nodes. After being served at node j, a task jumps to node k with probability p_{jk} and leaves the network with probability $p_j^* = 1 - \sum_k p_{jk}$. Upon arrival at any node, a task selects m of the N channels there at random and joins the one with the shortest queue. The state of the network at time $t \ge 0$ may be described by the vector $\underline{\mathbf{r}}(t) = \{r_j(n,t), 1 \le j \le J, n \in \mathbb{Z}_+\}$, where $r_j(n,t)$ is the proportion of channels at node j with queue length at least n at time t. We analyse the limit $N \to \infty$. We show that, under a standard non-overload condition, the limiting invariant distribution (ID) of the process $\underline{\mathbf{r}}$ is concentrated at a single point, and the process itself asymptotically approaches a single trajectory. This trajectory is identified with the solution to a countably infinite system of ODEs, whose fixed point corresponds to the limiting ID. Under the limiting ID, the tail of the distribution of queue-lengths decays super-exponentially, rather than exponentially as in the case of standard Jackson networks — hence the term "Fast Networks" in the title of the paper.

1 Introduction

The class of Jackson networks, introduced in [2], [3], remains one of the most popular and widely studied in queueing network theory. Attractions of a Jackson network lie in the simplicity of its construction, and in the partial exact solvability expressed by the product formula for the stationary distribution of the Markov process describing the evolution of the state of the network.

[1]Work supported by EPSRC studentship no. 95007341.

[2]Work supported in part by the EC grant 'Training Mobility and Research' (Contracts CHRX-CT 93-0411 and ERBMRXT-CT 960075A) and the INTAS grant 'Mathematical Methods for Stochastic Discrete Event Systems' (INTAS 93-820).

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

Key words and phrases: Jackson network, queueing network, dynamic routing, Markov process, super-exponential decay.

In [8], Vvedenskaya, Dobrushin and Karpelevich consider a model of a service station consisting of a large number of separate servers, and show that allowing tasks a small amount of flexibility about which queue to join produces considerable improvement in terms of average queue length. In this paper, we consider networks of Jackson type whose nodes are stations of this kind, and show that flexibility of routing leads to improvement in network performance in a similar way.

The model considered in [8] is as follows. The system consists of N identical channels, each with an infinite buffer and a single server with service rate μ. The input flow is Poisson with rate $N\lambda$; service times and arrival times are all independent. Upon arrival, each task selects m channels at random, and joins the one whose queue is the shortest. This model may be considered as a very simple example of *dynamic routing.* Similar problems are considered in [7] by Turner, who uses a coupling argument to provide stronger results comparing the evolution of the system for different values of m, and by Mitzenmacher in [6]. See also the recent paper [9], where a similar model with a more general class of routing regimes and service distributions is considered, and the review papers [5], where routing policies are discussed in the context of loss networks, and [4].

The state of the system may be described by the vector $\mathbf{r}_N = \{r_N(n), n \in \mathbb{Z}_+\}$, (here and below $\mathbb{Z}_+$ is the set of non-negative integers), where $r_N(n) = N^{-1}\sum_{n'\geq n} M(n')$, and $M(n')$ is the number of channels with queue length n'. Hence $r_N(n)$ is the proportion of queues in the system whose length is at least n. The process $\mathbf{r}_N(t) = \{r_n(n,t), n \in \mathbb{Z}_+, t \geq 0\}$ describing the state of the network at times $t \geq 0$ is easily seen to be a Markov process, with state space given by

$$\overline{\mathcal{U}}_N = \{\mathbf{g} = (g(n), n \in \mathbb{Z}_+) : g(0) = 1, g(n) \geq g(n+1) \geq 0, Ng(n) \in \mathbb{N} \text{ for all } n\} \tag{1.1}$$

If $\lambda < \mu$, the process is positive recurrent, with a unique stationary distribution π_N, and the main result of [8] is that

$$\lim_{N\to\infty} \mathbb{E}_{\pi_N} r_N(n) = \left(\frac{\lambda}{\mu}\right)^{(m^n-1)/(m-1)}, \ n \geq 0. \tag{1.2}$$

Thus, as $N \to \infty$, a "typical server" in the system will have $\geq n$ tasks in its buffer with a probability that decays super-exponentially as $n \to \infty$. We may compare this to a "linear" system, in which arriving tasks choose a channel at random; this system is equivalent to N isolated $M/M/1$ queues with arrival and service rates λ and μ. A typical server in the linear system has $\geq n$ tasks in the buffer with probability $(\lambda/\mu)^n$, $n \geq 0$, (independently of N), which is larger than the RHS of (1.2) and decays only exponentially as $n \to \infty$.

In fact, as is shown in [8], the whole process $\{\mathbf{r}_N(t)\}$ is asymptotically deterministic as $N \to \infty$. More precisely, extend (1.1) in the natural way by defining

$$\overline{\mathcal{U}} = \{\mathbf{g} = (g(n), n \in \mathbb{Z}_+) : g(0) = 1, g(n) \geq g(n+1) \geq 0 \text{ for all } n\}. \tag{1.3}$$

Then if the distribution of the initial state $\mathbf{r}_N(0)$ approaches the Dirac delta-measure concentrated at a point $\mathbf{g} \in \overline{\mathcal{U}}$, the distribution of $\{\mathbf{r}_N(t)\}$ is concentrated in the limit on the trajectory $\mathbf{u}(t, \mathbf{g}) = \{u(n, t, \mathbf{g}), n \in \mathbb{Z}_+\}$, $t \geq 0$, giving the unique solution to the following system of differential equations:

$$\dot{u}(n,t) = \lambda[u(n-1,t)^m - u(n,t)^m] - \mu[u(n,t) - u(n+1,t)],\ n \geq 1 \tag{1.4}$$

with boundary conditions

$$u(n,0) = g(n),\ n \geq 1, \quad \text{and} \quad u(0,t) = 1 \text{ for all } t.$$

The RHS of (1.2) provides a stationary solution to (1.4).

We now briefly describe some details of standard Jackson networks. We consider networks with finitely many nodes $1, \ldots, J$ (models with infinitely many nodes are also possible but will not be covered here). Each node j has a single server with an infinite buffer, and service rate μ_j. The network is fed by a family of independent Poisson flows of rate $\lambda_1, \ldots, \lambda_J$, arriving at the corresponding nodes. After being served at node j, a task joins the queue at node k with probability p_{jk}, and leaves the network with probability $p_j^* = 1 - \sum_k p_{jk}$; here we require $0 \leq p_{jk} \leq 1$ for all j, k, and $\sum_k p_{jk} \leq 1$ for all j. We will also assume throughout the paper that the $J \times J$ matrix $(\mathbf{I} - \mathbf{P})$ is invertible; then $(\mathbf{I} - \mathbf{P})^{-1} = \sum_{i=0}^{\infty} \mathbf{P}^i$. This condition guarantees that the expected total number of nodes visited by a particular task is finite.

The state of a Jackson network at time $t \geq 0$ can be described by the vector $\underline{n}(t) = (n_1(t), \ldots, n_J(t))$, where $n_j(t)$ is the length of the queue at node j at time t. (Throughout the paper, underlined symbols refer to J-component vectors). It is easy to see that $\underline{n}(t)$ is a Markov process, with state space $\mathbb{Z}_+^J$. Its stationary distribution, if one exists, may be determined as follows. Consider the equation

$$\underline{\rho} = \underline{\lambda} + \underline{\rho}\mathbf{P}. \tag{1.5}$$

Here, $\underline{\lambda} = (\lambda_1, \ldots, \lambda_J)$ is the vector of external arrival rates, and $\mathbf{P} = \{p_{jk}\}$ is the routing matrix. The vector $\underline{\rho} = (\rho_1, \ldots, \rho_J)$ is the unknown, and an entry ρ_j will represent the "effective arrival rate" at node j, including arrivals from inside as well as outside the network. (1.5) has solution

$$\underline{\rho} = \underline{\lambda}(\mathbf{I} - \mathbf{P})^{-1}. \tag{1.6}$$

Under the *non-overload condition*

$$\underline{\rho} < \underline{\mu}, \quad (\text{i.e. } \rho_j < \mu_j \text{ for all } j), \tag{1.7}$$

the network is positive recurrent, and its stationary distribution may be given by

$$\mathbb{P}(N_j \geq r_j \ \forall j) = \prod_j \left(\frac{\rho_j}{\mu_j}\right)^{r_j}, \quad \mathbf{r} = (r_1, \dots, r_J) \in \mathbb{Z}_+^J, \tag{1.8}$$

which is the product of the geometric distributions with paramters (ρ_j/μ_j) at node j, $1 \leq j \leq J$.

2 Main results

We modify the above model of a Jackson network in the following way. Each node $j \in \{1, \dots, J\}$ will now contain N channels, each of which has an infinite buffer and a single server with service rate μ_j. The network is fed by a family of independent Poisson flows of rates $N\lambda_1, \dots, N\lambda_J$ arriving at the corresponding nodes. As before, after being served at node j, a task jumps to node k with probability p_{jk} and leaves the network with probability p_j^*. Upon arrival at node j, either from outside or from inside the network, a task selects m of the N channels there uniformly at random (with replacement, though this becomes unimportant as N becomes large), and enters the one with the shortest queue (breaking ties at random if necessary). The network is therefore specified by the parameters J, $\underline{\lambda}$, $\underline{\mu}$, $\mathbf{P}$, N and m.

The state of the network at time $t \geq 0$ may be described by the collection $\underline{\mathbf{r}}(t) = \{r_j(n,t), 1 \leq j \leq J, n \in \mathbb{Z}_+\}$ where $r_j(n,t) = N^{-1} \sum_{n' \geq n} M_j(n',t)$ and $M_j(n',t)$ is the number of channels at node j with queue length n at time t. Hence $r_j(n)$ measures the proportion of queues at j whose length is n or greater. The process $\{\underline{\mathbf{r}}(t), t \geq 0\}$ (we will generally write $\{\underline{\mathbf{r}}_N(t)\}$ to stress the dependence on N) forms a Markov process, whose state space is the Cartesian product $\overline{\mathcal{U}}_N^J$, where $\overline{\mathcal{U}}_N$ is defined in (1.1). The generator of $\{\underline{\mathbf{r}}_N(t)\}$ is given by:

$$\begin{aligned}
\mathbf{A}_N f(\underline{\mathbf{g}}) =& N \sum_{n \geq 1} \sum_{1 \leq j \leq J} \lambda_j \left[g_j(n-1)^m - g_j(n)^m\right] \left[f(\underline{\mathbf{g}} + \frac{\underline{\mathbf{e}}_j(n)}{N}) - f(\underline{\mathbf{g}})\right] \\
&+ N \sum_{n \geq 1} \sum_{1 \leq j \leq J} \mu_j p_j^* \left[g_j(n) - g_j(n+1)\right] \left[f(\underline{\mathbf{g}} - \frac{\underline{\mathbf{e}}_j(n)}{N}) - f(\underline{\mathbf{g}})\right] \\
&+ N \sum_{n,n' \geq 1} \sum_{1 \leq j,k \leq J} \mu_k p_{kj} \left[g_k(n') - g_k(n'+1)\right] \\
&\times \left[\{g_j(n-1) - \frac{1}{N}\delta_{j,k}\delta_{n-1,n'}\}^m - \{g_j(n) - \frac{1}{N}\delta_{j,k}\delta_{n,n'}\}^m\right] \\
&\times \left[f(\underline{\mathbf{g}} + \frac{\underline{\mathbf{e}}_j(n)}{N} - \frac{\underline{\mathbf{e}}_k(n')}{N}) - f(\underline{\mathbf{g}})\right],
\end{aligned} \tag{2.1}$$

where $\underline{\mathbf{g}} = \{g_j(n), n \in \mathbb{Z}_+, 1 \leq j \leq J\} \in \overline{\mathcal{U}}_N^J$ and f is a function from $\overline{\mathcal{U}}_N^J$ to $\mathbb{R}$; here $\underline{\mathbf{e}}_k(n')$ denotes the vector $(e_j(n), 1 \leq j \leq J, n \in \mathbb{Z}_+)$ whose only non-zero entry is $e_k(n') = 1$, and $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise. We denote by $\mathbf{T}_N(t)$ the semigroup of transition operators generated by $\mathbf{A}_N$, defined formally by $\mathbf{T}_N(t) = \exp(t\mathbf{A}), t \geq 0$, and acting on functions $f : \overline{\mathcal{U}}_N^J \to \mathbb{R}$.

As well as the space $\overline{\mathcal{U}}^J$, where $\overline{\mathcal{U}}$ is defined in (1.3), we will also use the space $\mathcal{U}^J$, where

$$\mathcal{U} = \{g = (g(n), n \in \mathbb{Z}_+) : g(0) = 1, g(n) \geq g(n+1) \text{ for all } n, \sum_{n=0}^{\infty} g(n) < \infty\}. \tag{2.2}$$

The space $\mathcal{U}^J$ represents limiting states where the average queue length per channel is finite at each node. We will use the norm

$$\|\underline{\mathbf{u}}\| = \sup_{1 \leq j \leq J} \sup_{n \in \mathbb{Z}_+} \frac{|u_j(n)|}{n+1} \tag{2.3}$$

on the spaces $\mathcal{U}^J, \overline{\mathcal{U}}^J$ and $\overline{\mathcal{U}}_N^J$; this norm is understood when we refer to continuity and convergence below. Under this norm $\overline{\mathcal{U}}^J$ becomes a complete compact metric space. Note that $\overline{\mathcal{U}}$ may be interpreted as the space of sub-probability measures on $\mathbb{Z}_+$; and $\mathcal{U}$ as the subspace of probability measures with a finite first moment; the topology generated by norm (2.3) then corresponds to the topology of weak convergence on the one-point compactification $(\mathbb{Z}_+ \cup \{\infty\})$ of $\mathbb{Z}_+$. The choice of these spaces leads to a convenient description in the limit $N \to \infty$.

Our main results are stated in Theorems 1–4 below. We will frequently refer to the non-overload condition (1.7), whose form is unchanged in the modified setting. An important role will be played by the following infinite system of non-linear differential equations (corresponding to the system (1.4) used in [8]) for $\underline{\mathbf{u}}(t) = \{u_j(n,t), 1 \leq j \leq J, n \in \mathbb{Z}_+\}$, $t \geq 0$, with initial condition $\underline{\mathbf{g}} \in \overline{\mathcal{U}}^J$:

$$\underline{\mathbf{u}}(0) = \underline{\mathbf{g}}, \tag{2.4}$$

$$\dot{\underline{\mathbf{u}}}(t) = \underline{\mathbf{h}}(\underline{\mathbf{u}}(t)), \tag{2.5}$$

where, for all j,

$$h_j(0, \underline{\mathbf{u}}) = 0, \tag{2.6}$$

$$h_j(n, \underline{\mathbf{u}}) = \left[\lambda_j + \sum_{1 \leq k \leq J} \mu_k p_{kj} u_k(1)\right] [u_j(n-1)^m - u_j(n)^m] - \mu_j [u_j(n) - u_j(n+1)] \tag{2.7}$$

for all $n \geq 1$.

The first theorem establishes various properties of this system, and the second theorem describes the way in which the behaviour of the processes $\underline{\mathbf{r}}_N(t)$ converges asymptotically to that of the limiting system of ODEs as $N \to \infty$, for finite t.

Theorem 1
(i) If $\underline{\mathbf{g}} \in \overline{\mathcal{U}}^J$, the system (2.4)–(2.7) has a unique solution $\underline{\mathbf{u}}(t, \underline{\mathbf{g}}), t \geq 0$ in $\overline{\mathcal{U}}^J$.
(ii) Under (1.7) there exists a unique fixed point $\underline{\mathbf{a}} \in \mathcal{U}^J$ such that $\underline{\mathbf{u}}(t, \underline{\mathbf{a}}) = \underline{\mathbf{a}}$ for all t, with

$$a_j(n) = \left(\frac{\rho_j}{\mu_j}\right)^{(m^n-1)/(m-1)}. \tag{2.8}$$

(iii) Under (1.7)

$$\lim_{t \to \infty} \underline{\mathbf{u}}(t, \underline{\mathbf{g}}) = \underline{\mathbf{a}} \quad \text{for all } \underline{\mathbf{g}} \in \mathcal{U}^J. \tag{2.9}$$

Thus there exists a unique probability measure π on $\mathcal{U}^J$ which is invariant under the map $\underline{\mathbf{g}} \mapsto \underline{\mathbf{u}}(t, \underline{\mathbf{g}})$, so that

$$\int f(\underline{\mathbf{g}}) d\pi(\underline{\mathbf{g}}) = \int f(\underline{\mathbf{u}}(t, \underline{\mathbf{g}})) d\pi(\underline{\mathbf{g}})$$

for all $t \geq 0$, $f : \overline{\mathcal{U}}^J \to \mathbb{R}$; and $\pi = \delta_{\underline{\mathbf{a}}}$, the probability measure concentrated at the fixed point $\underline{\mathbf{a}}$.

Theorem 2 *For any continuous function $f : \overline{\mathcal{U}}^J \to \mathbb{R}$ and $t \geq 0$,*

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

and the convergence is uniform in t within any bounded interval.

The next result provides a coupling which compares the behaviour of the network described by the process $\underline{\mathbf{r}}_N(t)$ with that of a "linear network" where $m = 1$. See the remark after the proof of Theorem 3 for suggestions regarding an extension of the theorem.

Theorem 3 *Let Q and R be two networks with the same parameters $J, \underline{\lambda}, \underline{\mu}, \mathbf{P}$ and N, and the same state at $t = 0$, and with $m(Q) = 1$, $m(R) \geq 1$. Then there is a coupling of the processes $\{\underline{\mathbf{r}}_N^{(Q)}(t)\}$ and $\{\underline{\mathbf{r}}_N^{(R)}(t)\}$ such that, for all $t \geq 0$, the total number of tasks present in R at time t is no greater than the total number in Q at t.*

The final result demonstrates the convergence of the stationary distributions, where they exist, to the probability measure concentrated at the fixed point of the system of ODEs. Thus, from the RHS of (2.8), we have, for $m > 1$, super-exponential decay of the tail of the queue length distribution in the limiting stationary regime, in contrast to the exponential decay occurring in the case $m = 1$. In other words, the networks which we consider have much shorter queues (per channel) than "standard" Jackson networks. This is the reason for the term "Fast Networks" in the title of the paper.

Theorem 4 *Under the non-overload condition (1.7),*
(i) The Markov process $\underline{\mathbf{r}}_N(t)$ is positive recurrent for all N, and hence has a unique invariant distribution π_N for each N.
(ii) $\pi_N \to \delta_{\underline{\mathbf{a}}}$ weakly, where $\delta_{\underline{\mathbf{a}}}$ is given by Theorem 1,(ii) and (iii), i.e.

$$\lim_{N\to\infty} \mathbb{E}_{\pi_N} f(\underline{\mathbf{g}}) = f(\underline{\mathbf{a}})$$

for all continuous functions $f : \overline{\mathcal{U}}^J \to \mathbb{R}$.

The rest of the paper is devoted to the proof of Theorems 1–4. The methods used to prove Theorems 1 and 2 are essentially those that may be used in the non-network case, with some modifications; we are also able to simplify in some aspects the approach used in [8]. Later results, especially the proof of Theorem 3, require more particular attention to the network structure.

3 Limiting equations and convergence for finite t

Proof of Theorem 1(i): Define $\xi(x) = [\min(x,1)]_+$, where $[y]_+ = \max(y,0)$, and consider the following modification of (2.4)–(2.7):

$$\underline{\mathbf{u}}(0) = \underline{\mathbf{g}}, \tag{2.4$'$}$$

$$\dot{\underline{\mathbf{u}}}(t) = \tilde{\underline{\mathbf{h}}}(\underline{\mathbf{u}}(t)), \tag{2.5$'$}$$

where, for all j.

$$\tilde{h}_j(0,\underline{\mathbf{u}}) = 0, \tag{2.6$'$}$$

$$\begin{aligned}\tilde{h}_j(n,\underline{\mathbf{u}}) = &\left[\lambda_j + \sum_{1\le k\le J} \mu_k p_{kj}\xi(u_k(1))\right] \times [\xi(u_j(n-1))^m - \xi(u_j(n))^m]_+ \\ &- \mu_j[\xi(u_j(n)) - \xi(u_j(n+1))]_+ \quad \text{for all } n \ge 1.\end{aligned} \tag{2.7$'$}$$

Since the RHS of (2.7) and (2.7′) are the same if $\underline{\mathbf{u}} \in \overline{\mathcal{U}}^J$, the system (2.4)–(2.7) has the same solutions in $\overline{\mathcal{U}}^J$ as the system (2.4′)–(2.7′). Also, if $\underline{\mathbf{g}} \in \overline{\mathcal{U}}^J$, then any solution of (2.4′)–(2.7′) remains within $\overline{\mathcal{U}}^J$, since, under (2.7′), if $u_j(n,t) \le u_j(n+1,t)$ for some j, n, t, then $\tilde{h}_j(n, \underline{\mathbf{u}}(t)) \ge 0$ and $\tilde{h}_j(n+1, \underline{\mathbf{u}}(t)) \le 0$, and if $u_j(n,t) \le 0$ for some j, n, t, then $\tilde{h}_j(n, \underline{\mathbf{u}}) \ge 0$. Thus, in order to show that there exists a unique solution to (2.4)–(2.7) in $\overline{\mathcal{U}}^J$, it suffices to show that there exists a unique solution to (2.4′)–(2.7′) in $(\mathbb{R}^{\mathbb{Z}_+})^J$. We use the Picard successive approximation method. The norm (2.3) is now extended to the whole space $(\mathbb{R}^{\mathbb{Z}_+})^J$ (allowing the value ∞ where required). From (2.7′), for all $\underline{\mathbf{u}}, \underline{\mathbf{u}}' \in (\mathbb{R}^{\mathbb{Z}_+})^J$,

$$\begin{aligned} \|\tilde{\underline{\mathbf{h}}}(\underline{\mathbf{u}})\| &\le \max_j \lambda_j + (J+1) \max_j \mu_j, \\ &:= C_1, \end{aligned} \tag{3.1}$$

and

$$\begin{aligned} \|\tilde{\underline{\mathbf{h}}}(\underline{\mathbf{u}}) - \tilde{\underline{\mathbf{h}}}(\underline{\mathbf{u}}')\| &\le (2m \max_j \lambda_j + (6+2m) \max_j \mu_j)\|\underline{\mathbf{u}} - \underline{\mathbf{u}}'\| \\ &:= C_2\|\underline{\mathbf{u}} - \underline{\mathbf{u}}'\|. \end{aligned} \tag{3.2}$$

Here we have used the facts that $|\xi(u) - \xi(u')| \le |u - u'|$, and, for $a_1, a_2, b_1, b_2 \in [0,1]$,

$$|a_1 b_1^m - a_2 b_2^m| \le |a_1 - a_2| + m|b_1 - b_2|.$$

For $t \ge 0$, let $\underline{\mathbf{u}}^{(0)}(t) = \underline{\mathbf{g}}$, and, recursively, let

$$\underline{\mathbf{u}}^{(r)}(t) = \underline{\mathbf{g}} + \int_0^t \tilde{\underline{\mathbf{h}}}(\underline{\mathbf{u}}^{(r-1)}(s))ds, \quad r \in \mathbb{N}. \tag{3.3}$$

It follows by induction that $\underline{\mathbf{u}}^{(r)}(t)$ is continuous in t on $[0,\infty)$ for all r, and that

$$\|\underline{\mathbf{u}}^{(r+1)}(t) - \underline{\mathbf{u}}^{(r)}(t)\| \le \frac{C_1 C_2^r t^{r+1}}{(r+1)!} \quad \text{for all } r, t.$$

Hence, for all $t \ge 0$, $\underline{\mathbf{u}}(s) = \lim_{r\to\infty} \underline{\mathbf{u}}^{(r)}(s)$ exists uniformly for $s \in [0,t]$. Since $\tilde{\underline{\mathbf{h}}}$ is uniformly continuous (see (3.2)), (3.3) then implies that

$$\underline{\mathbf{u}}(t) = \underline{\mathbf{g}} + \int_0^t \tilde{\underline{\mathbf{h}}}(\underline{\mathbf{u}}(s))ds, \quad t \ge 0,$$

and hence, by differentiating, that $\underline{\mathbf{u}}(t)$ is a solution to (2.4′)–(2.7′).

To show uniqueness, suppose that $\underline{\mathbf{w}}(t)$ is any solution of (2.4′)–(2.7′). Then

$$\underline{\mathbf{w}}(t) = \underline{\mathbf{g}} + \int_0^t \tilde{\underline{\mathbf{h}}}(\underline{\mathbf{w}}(s))ds, \tag{3.4}$$

and so

$$\underline{\mathbf{w}}(t) - \underline{\mathbf{u}}^{(r)}(t) = \int_0^t [\tilde{\underline{\mathbf{h}}}(\underline{\mathbf{w}}(s)) - \tilde{\underline{\mathbf{h}}}(\underline{\mathbf{u}}^{(r-1)}(s))]ds. \tag{3.5}$$

It follows by induction, as before, that

$$\|\underline{\mathbf{w}}(t) - \underline{\mathbf{u}}^{(r)}(t)\| \leq \frac{C_1 C_2^r t^{r+1}}{(r+1)!},$$

so that $\underline{\mathbf{w}}(t) = \lim_{r\to\infty} \underline{\mathbf{u}}^{(r)}(t) = \underline{\mathbf{u}}(t)$. This completes the proof of Theorem 1(i). $\square$

Before proving the rest of Theorem 1, we turn to the proof of Theorem 2, which provides a result which we will use in showing the convergence in Theorem 1(iii). The next two lemmas show the existence of the first and second derivatives of $\underline{\mathbf{u}}(t, \underline{\mathbf{g}})$ with respect to the entries in $\underline{\mathbf{g}}$, and give bounds on their size. Lemma 3.1 is Proposition 1 from [8]; we include the brief proof for the sake of completeness.

Lemma 3.1 *Consider the infinite system of differential equations*

$$\begin{aligned} z_x(0) &= c_x \\ \frac{dz_x(t)}{dt} &= \sum_{y\in\mathcal{C}} a_{x,y}(t) z_y(t) + b_x(t), \end{aligned}$$

for $x \in \mathcal{C}$, $t \geq 0$, where $\mathcal{C}$ is a countable index set and where, for all $x, y \in \mathcal{C}$, the functions $a_{x,y}(t)$ and $b_x(t)$ are continuous in t, and suppose that $\sum_{y\in\mathcal{C}} |a_{x,y}(t)| \leq a$, $|b_x(t)| \leq b_0 e^{bt}$, $|c_x| \leq c$ for all x, t. Then there exists a unique solution $\{z_x(t)\}$, $x \in \mathcal{C}$, $t \geq 0$, and

$$|z_x(t)| \leq ce^{at} + \frac{b_0}{b-a}(e^{bt} - e^{at}) \text{ for all } x, t. \tag{3.6}$$

Proof: Since the system is countable and linear with coefficients bounded uniformly in x and in t from any bounded interval, one may show, for example using a successive approximation argument similar to that in the proof of Theorem 1(i) above, that it has a unique solution. To show the bound (3.6) we again use the successive approximation method. Let

$$z_x^{(0)}(t) = c_x + \int_0^t b_x(s)ds,$$

and

$$z_x^{(k+1)} = c_x + \int_0^t \left[\sum_{y\in\mathcal{C}} a_{x,y}(s) z_y^{(k)}(s) + b_x(s)\right] ds, \quad \text{for } k \in \mathbb{Z}_+.$$

Then

$$|z_x^{(0)}(t)| \leq c + \frac{b_0}{b}(e^{bt} - 1),$$

and an induction gives

$$|z_x^{(k)}(t) - z_x^{(k-1)}(t)| \leq a^k \left(\frac{ct^k}{k!} + \frac{b_0}{b^{k+1}} \sum_{l=k+1}^{\infty} \frac{(bt)^l}{l!} \right), \quad \text{for } x \in \mathcal{C}, k \geq 1.$$

Taking the limit $k \to \infty$ and summing over k gives the required result. $\square$

Lemma 3.2 *The derivatives* $\dfrac{\partial \underline{\mathbf{u}}(t, \underline{\mathbf{g}})}{\partial g_j(n)}$, $\dfrac{\partial^2 \underline{\mathbf{u}}(t, \underline{\mathbf{g}})}{\partial g_j(n)^2}$ *and* $\dfrac{\partial^2 \underline{\mathbf{u}}(t, \underline{\mathbf{g}})}{\partial g_j(n) \partial g_{j'}(n')}$ *exist, for* $\underline{\mathbf{g}} \in \overline{\mathcal{U}}^J$ *and* $t \geq 0$*, and satisfy*

$$\left| \frac{\partial u_k(r, t, \underline{\mathbf{g}})}{\partial g_j(n)} \right| \leq e^{A_1 t}$$

$$\left| \frac{\partial^2 u_k(r, t, \underline{\mathbf{g}})}{\partial g_j(n)^2} \right|, \quad \left| \frac{\partial^2 u_k(r, t, \underline{\mathbf{g}})}{\partial g_j(n) \partial g_{j'}(n')} \right| \leq \frac{A_2}{A_1}(e^{2A_1 t} - e^{A_1 t}),$$

where

$$\begin{aligned} A_1 &= 2(Jm + J + 1) \max_j \mu_j + 2m \max_j \lambda_j, \\ A_2 &= 2J(m(m-1) + 2) \max_j \mu_j + 2m(m-1) \max_j \lambda_j. \end{aligned}$$

Proof: Fix $j, n, \underline{\mathbf{g}}$ and write $\underline{\mathbf{u}}'(t) = \partial \underline{\mathbf{u}}(t, \underline{\mathbf{g}}) / \partial g_j(n)$. If this derivative exists, then $\underline{\mathbf{u}}'(t)$ satisfies $u_k'(0, t) = 0$, $u_k'(r, 0) = \delta_{k,j} \delta_{r,n}$, and, by differentiating (2.7), (we omit the argument t to simplify the notation):

$$\begin{aligned} \frac{du_k'(r)}{dt} &= \left[\lambda_j + \sum_{1 \leq i \leq J} \mu_i p_{ik} u_i(1) \right] [\mu u_k(r-1)^{m-1} u_k'(r-1) + \mu u_k(r)^{m-1} u_k'(r)] \\ &\quad + \left[\sum_{1 \leq i \leq J} \mu_i p_{ik} u_i'(1) \right] [u_k(r-1)^m - u_k(r)^m] - \mu_k [u_k'(r) - u_k'(r+1)]. \end{aligned}$$

Conversely if $\underline{\mathbf{u}}'(t)$ is a solution of this system, then it works as the required derivative. Using the fact that $|u_k(r)| \leq 1$ for all k, r, we may now apply Lemma 3.1 with $a = A_1$, $b_0 = 0$, $c = 1$ and $\mathcal{C} = \mathbb{Z}_+ \times \{1, \ldots, J\}$ to give the first bound.

By differentiating the above system again with respect to $g_j(n)$ or $g_{j'}(n')$, one arrives at systems of equations for the required second partial derivatives, with coefficients now involving $\underline{\mathbf{u}}$ and $\underline{\mathbf{u}}'$, to which Lemma 3.1 can be applied with $a = A_1$, $b = 2A_1$, $b_0 = A_2$, $c = 0$ to give the second bound. $\square$

Proof of Theorem 2: Let L be the set of continuous functions $f : \overline{\mathcal{U}}^J \to \mathbb{R}$, and let D be the set of those $f \in L$ for which the derivatives

$$\frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)}, \frac{\partial^2 f(\underline{\mathbf{g}})}{\partial g_j(n)^2} \quad \text{and} \quad \frac{\partial^2 f(\underline{\mathbf{g}})}{\partial g_j(n) \partial g_{j'}(n')}$$

exist for all $\underline{\mathbf{g}}, j, j', n, n'$ and are uniformly bounded in modulus by some constant $C = C(f) < \infty$. Observe that D is dense in L, using the norm (2.3) on $\overline{\mathcal{U}}^J$ and the sup norm on L. For $f \in D$,

$$N\left(f\left(\underline{\mathbf{g}} + \frac{\underline{\mathbf{e}}_j(n)}{N}\right) - f(\underline{\mathbf{g}})\right) \to \frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)}$$
$$N\left(f\left(\underline{\mathbf{g}} - \frac{\underline{\mathbf{e}}_j(n)}{N}\right) - f(\underline{\mathbf{g}})\right) \to -\frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)}$$

and

$$N\left(f\left(\underline{\mathbf{g}} + \frac{\underline{\mathbf{e}}_j(n)}{N} - \frac{\underline{\mathbf{e}}_k(n')}{N}\right) - f(\underline{\mathbf{g}})\right) \to \frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)} - \frac{\partial f(\underline{\mathbf{g}})}{\partial g_k(n')}$$

uniformly in $\underline{\mathbf{g}}$ from $\overline{\mathcal{U}}^J$. Thus, using (2.1) we have

$$\begin{aligned} A_N f(\underline{\mathbf{g}}) \to & \sum_{n\geq 1} \sum_{1\leq j\leq J} \lambda_j [g_j(n-1)^m - g_j(n)^m] \frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)} \\ & - \sum_{n\geq 1} \sum_{1\leq j\leq J} \mu_j p_j^* [g_j(n) - g_j(n+1)] \frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)} \\ & + \sum_{n,n'\geq 1} \sum_{1\leq j,k\leq J} \mu_k p_{kj} [g_k(n') - g_k(n'+1)][g_j(n-1)^m - g_j(n)^m] \\ & \quad \times \left[\frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)} - \frac{\partial f(\underline{\mathbf{g}})}{\partial g_k(n')}\right], \end{aligned} \tag{3.7}$$

uniformly in $\underline{\mathbf{g}}$. The RHS of (3.7) may be rewritten as

$$\sum_{n\geq 1} \sum_{1\leq j\leq J} \frac{\partial f(\underline{\mathbf{g}})}{\partial g_j(n)} \left\{ \left[\lambda_j + \sum_{1\leq k\leq J} \mu_k p_{kj} g_k(1)\right] [g_j(n-1)^m - g_j(n)^m] - \mu_j [g_j(n) - g_j(n+1)] \right\}$$

which coincides with

$$\frac{d}{dt} f(\underline{\mathbf{u}}(t, \underline{\mathbf{g}}))\Big|_{t=0}, \tag{3.8}$$

where $\underline{\mathbf{u}}(t, \underline{\mathbf{g}})$ is the solution to (2.4)–(2.5).

Setting $\mathbf{T}(t)f(\underline{\mathbf{g}}) = f(\underline{\mathbf{u}}(t,\underline{\mathbf{g}}))$ defines a semigroup of operators $\mathbf{T}(t), t \geq 0$ in L, corresponding to shifts along the solutions of (2.4)–(2.5). The generator $\mathbf{A}$ of this semigroup is given by (3.8). Thus we have

$$\lim_{N\to\infty} \mathbf{A}_N f = \mathbf{A} f \tag{3.9}$$

in the sup norm, for all $f \in D$.

Define $D_0 \subset D$ as the set of those functions in D which depend on only finitely many variables $g_j(n)$. By the definition of the norm (2.3) on $\overline{\mathcal{U}}^J$, D_0 is dense in D, and hence in L. Also, it follows from Lemma 3.2 that $\mathbf{T}(f_0) \in D$ for all $f_0 \in D_0$. Finally, observe that the semigroups $\mathbf{T}_N(t)$ and $\mathbf{T}(t)$ are continuous and contracting in L. These facts, together with (3.9) and with Proposition 3.3 and Theorem 6.1 from Chapter 1 of [1], give the result of Theorem 2. $\square$

We write $\underline{\mathbf{g}} \leq \underline{\mathbf{g}}'$ to mean that all the inequalities $g_j(n) \leq g'_j(n)$ hold; similarly $\min[\underline{\mathbf{g}}, \underline{\mathbf{g}}']$ and $\max[\underline{\mathbf{g}}, \underline{\mathbf{g}}']$ are defined componentwise.

Lemma 3.3 *If $\underline{\mathbf{g}} \leq \underline{\mathbf{g}}'$, where $\underline{\mathbf{g}}, \underline{\mathbf{g}}' \in \overline{\mathcal{U}}^J$, then $\underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \leq \underline{\mathbf{u}}(t,\underline{\mathbf{g}}')$ for all t.*

Proof: Define the coordinate functions $f_{j,n}(\underline{\mathbf{g}}) = g_j(n), n \in \mathbb{Z}_+, 1 \leq j \leq J$. Fix N, and consider two networks with initial states $\underline{\mathbf{g}}_N, \underline{\mathbf{g}}'_N \in \mathcal{U}^J_N$, $\underline{\mathbf{g}}_N \leq \underline{\mathbf{g}}'_N$. A simple coupling shows that

$$\mathbf{T}_N(t) f_{j,n}(\underline{\mathbf{g}}_N) \leq \mathbf{T}_N(t) f_{j,n}(\underline{\mathbf{g}}'_N), \quad \text{for all} \quad t \geq 0, j, n. \tag{3.10}$$

Now for given $\underline{\mathbf{g}}, \underline{\mathbf{g}}' \in \overline{\mathcal{U}}^J$ with $\underline{\mathbf{g}} \leq \underline{\mathbf{g}}'$, choose sequences $\{\underline{\mathbf{g}}_N\}, \{\underline{\mathbf{g}}'_N\}, N \in \mathbb{N}$ with $\underline{\mathbf{g}}_N, \underline{\mathbf{g}}'_N \in \mathcal{U}^J_N$, such that $\underline{\mathbf{g}}_N \leq \underline{\mathbf{g}}'_N$ for all N and $\underline{\mathbf{g}}_N \to \underline{\mathbf{g}}$, $\underline{\mathbf{g}}'_N \to \underline{\mathbf{g}}'$ as $N \to \infty$. Then it follows from Theorem 2 and (3.10) that $f_{j,n}(\underline{\mathbf{u}}(t,\underline{\mathbf{g}})) \leq f_{j,n}(\underline{\mathbf{u}}(t,\underline{\mathbf{g}}'))$ for all t, j, n, which is exactly what we need. $\square$

We define the quantities $v_j(n,\underline{\mathbf{g}}) = \sum_{n' \geq n} g_j(n')$, $n \geq 1$, $1 \leq j \leq J$, $\underline{\mathbf{g}} \in \mathcal{U}^J$. Observe that $v_j(1,\underline{\mathbf{g}})$ represents the average queue length of the channels at node j in the state $\underline{\mathbf{g}}$, which is guaranteed to be finite by the definition of $\mathcal{U}$ (2.2); also note that $v_j(n,\underline{\mathbf{g}}) \geq v_j(n+1,\underline{\mathbf{g}})$ for all j, n. In addition, given a solution $\underline{\mathbf{u}}(t,\underline{\mathbf{g}})$ of the system (2.4)–(2.7), write $v_j(n,t,\underline{\mathbf{g}}) = v_j(n,\underline{\mathbf{u}}(t,\underline{\mathbf{g}}))$. Define the vectors $\underline{v}(n,t,\underline{\mathbf{g}}) = (v_1(n,t,\underline{\mathbf{g}}), \ldots, v_J(n,t,\underline{\mathbf{g}}))$, and similarly $\underline{u}(n,t,\underline{\mathbf{g}})$, $\underline{\dot{u}}(n,t,\underline{\mathbf{g}})$, etc., and write $\underline{x} \cdot \underline{y} = (x_1 y_1, \ldots, x_J y_J)$.

Lemma 3.4 *If $\underline{\mathbf{g}} \in \mathcal{U}^J$ then $\underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \in \mathcal{U}^J$ for all $t \geq 0$, and*

$$\dot{v}_j(n,t,\underline{\mathbf{g}}) = \left[\lambda_j + \sum_{1 \leq k \leq J} \mu_k p_{kj} u_k(1,t,\underline{\mathbf{g}})\right] u_j(n-1,t,\underline{\mathbf{g}})^m - \mu_j u_j(n,t,\underline{\mathbf{g}}).$$

In particular,

$$\dot{v}_j(1,t,\underline{\mathbf{g}}) = \lambda_j + \sum_{1\le k\le J} \mu_k p_{kj} u_k(1,t,\underline{\mathbf{g}}) - \mu_j u_j(1,t,\underline{\mathbf{g}})$$

or, in vector notation,

$$\underline{\dot{v}}(1,t,\underline{\mathbf{g}})(\mathbf{I}-\mathbf{P})^{-1} = \underline{\rho} - \underline{\mu}\cdot\underline{u}(1,t,\underline{\mathbf{g}}). \tag{3.11}$$

Proof: Summing the RHS of (2.7) over $n' \ge n$ gives the result. □

Proof of Theorem 1, (ii) and (iii): For a fixed point $\underline{\mathbf{a}}$ of the map $\underline{\mathbf{u}}$, we need $\underline{\mathbf{h}}(\underline{\mathbf{a}}) = 0$ in (2.6) and (2.7). Restricting ourselves to the space $\mathcal{U}^J$, we have $a_j(0) = 1$ for all j, and can use (3.11), setting the LHS to 0, to give $a_j(1) = \rho_j/\mu_j$ for all j. Putting these values back into the RHS of (2.7) gives

$$a_j(n+1) = a_j(n) - \frac{\rho_j}{\mu_j}[a_j(n-1)^m - a_j(n)^m],$$

which can be solved recursively to confirm the unique solution $\underline{\mathbf{a}}$ given in (2.8).

To prove the convergence in (2.9), it will be sufficient to show that the conclusion $\underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \to \underline{\mathbf{a}}$ holds for all those $\underline{\mathbf{g}} \in \mathcal{U}^J$ for which either $\underline{\mathbf{g}} \le \underline{\mathbf{a}}$ or $\underline{\mathbf{g}} \ge \underline{\mathbf{a}}$, since Lemma 3.3 implies that

$$\underline{\mathbf{u}}(t,\min[\underline{\mathbf{g}},\underline{\mathbf{a}}]) \le \underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \le \underline{\mathbf{u}}(t,\max[\underline{\mathbf{g}},\underline{\mathbf{a}}]) \quad \text{for all } \underline{\mathbf{g}} \in \overline{\mathcal{U}}^J, t \ge 0.$$

First, we need to check that for such a $\underline{\mathbf{g}}$, the quantities $v_j(1,t,\underline{\mathbf{g}})$ (and hence also $v_j(n,t,\underline{\mathbf{g}})$, $n > 1$) remain bounded uniformly in t. If $\underline{\mathbf{g}} \le \underline{\mathbf{a}}$ then, by Lemma 3.3, $\underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \le \underline{\mathbf{a}}$, and so $\underline{\mathbf{v}}(1,t,\underline{\mathbf{g}}) \le \underline{\mathbf{v}}(1,\underline{\mathbf{a}})$, for all t.

On the other hand, if $\underline{\mathbf{g}} \ge \underline{\mathbf{a}}$, then by the same lemma, $\underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \ge \underline{\mathbf{a}}$. In particular, $u_j(1,t,\underline{\mathbf{g}}) \ge (\rho_j/\mu_j)$ for all j, or, in vector notation, $\underline{\mu}\cdot\underline{u}(1,t,\underline{\mathbf{g}}) \ge \underline{\rho}$. Returning to (3.11) yields

$$\underline{\dot{v}}(1,t,\underline{\mathbf{g}})(\mathbf{I}-\mathbf{P})^{-1} \le \underline{0}. \tag{3.12}$$

Since $(\mathbf{I}-\mathbf{P})^{-1}$ has diagonal entries ≥ 1 and all other entries ≥ 0, (3.12) implies that the entries of vectors $\underline{v}(1,t,\underline{\mathbf{g}})(\mathbf{I}-\mathbf{P})^{-1}$ and $\underline{v}(1,t,\underline{\mathbf{g}})$ remain bounded uniformly in t.

Since the derivative of $u_j(n,s,\underline{\mathbf{g}})$ is bounded for all j, the convergence $\underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \to \underline{\mathbf{a}}$ will follow from

$$\int_0^\infty [u_j(n,s,\underline{\mathbf{g}}) - a_j(n)]ds < \infty, \quad 1 \le j \le J, \quad n \in \mathbb{Z}_+ \tag{3.13}$$

in the case $\underline{\mathbf{g}} \ge \underline{\mathbf{a}}$, and from

$$\int_0^\infty [a_j(n) - u_j(n,s,\underline{\mathbf{g}})]ds < \infty, \quad 1 \le j \le J, \quad n \in \mathbb{Z}_+$$

in the case $\underline{\mathbf{g}} \leq \underline{\mathbf{a}}$; the integrands are non-negative for all s in each case. The two bounds may be proved similarly, and we discuss, say, (3.13). We use induction in n, starting with $n = 1$. Using (3.11),

$$\begin{aligned}\int_0^t [u_j(n,s,\underline{\mathbf{g}}) - a_j(n)]ds &= \frac{1}{\mu_j}\int_0^t ds[\underline{\mu}\cdot\underline{u}(1,s,\underline{\mathbf{g}}) - \underline{\rho}]_j \\ &= \frac{1}{\mu_j}\int_0^t ds\left[-\underline{\dot{v}}(1,s,\underline{\mathbf{g}})(\mathbf{I}-\mathbf{P})^{-1}\right]_j \\ &= \frac{1}{\mu_j}\left[(\underline{v}(1,0,\underline{\mathbf{g}}) - \underline{v}(1,t,\underline{\mathbf{g}}))(\mathbf{I}-\mathbf{P})^{-1}\right]_j\end{aligned}$$

The RHS remains bounded as $t \to \infty$, so the integral on the LHS converges.

Now assume that the integral (3.13) converges for all $n \leq L-1$. Using Lemma 3.4 and the relation

$$\mu_j a_j(L) = \left[\lambda_j + \sum_{1\leq k\leq J}\mu_k p_{kj} a_k(1)\right] a_j(L-1)^m,$$

we have

$$\begin{aligned}v_j(L,0,\underline{\mathbf{g}}) - v_j(L,t,\underline{\mathbf{g}}) &= -\int_0^t \dot{v}_j(L,s,\underline{\mathbf{g}})ds \\ &= \int_0^t\left[\mu_j u_j(L,s,\underline{\mathbf{g}}) - \left(\lambda_j + \sum_{1\leq k\leq J}\mu_k p_{kj} u_k(1,s,\underline{\mathbf{g}})\right) u_j(L-1,s,\underline{\mathbf{g}})^m\right] ds \\ &= \mu_j\int_0^t [u_j(L,s,\underline{\mathbf{g}}) - a_j(L)]ds - \lambda_j\int_0^t [u_j(L-1,s,\underline{\mathbf{g}})^m - a_j(L-1)^m]ds \\ &\quad - \sum_{1\leq k\leq J}\mu_k p_{kj}\int_0^t [u_k(1,s,\underline{\mathbf{g}})u_j(L-1,s,\underline{\mathbf{g}})^m - a_k(1)a_j(L-1)^m]ds\end{aligned}$$

By the induction hypothesis, the last two integrals converge as $t \to \infty$. The LHS also remains bounded, so we have $\int_0^\infty [u_j(L,s,\underline{\mathbf{g}}) - a_j(L)]ds < \infty$ as required.

The second statement in part (iii) of the theorem follows immediately from the convergence properties in the space $\mathcal{U}^J$ just established. $\square$

Remark: There are many other invariant distributions which are not concentrated on $\mathcal{U}^J$. In fact, for all $\underline{\epsilon}$ in a neighbourhood of $\underline{0}$ in $[0,1]^J$, there exists a fixed point $\underline{\mathbf{a}}^{(\underline{\epsilon})}$ of the maps $\underline{\mathbf{u}} : \underline{\mathbf{g}} \to \underline{\mathbf{u}}(t,\underline{\mathbf{g}})$ with $\lim_{n\to\infty} a_j^{(\underline{\epsilon})}(n) = \epsilon_j$ for all j, and if $\underline{\mathbf{g}} \in \overline{\mathcal{U}}^J$ satisfies $\lim_{n\to\infty} g_j(n) = \epsilon_j$ for all j, then $\underline{\mathbf{u}}(t,\underline{\mathbf{g}}) \to \underline{\mathbf{a}}^{(\underline{\epsilon})}$ as $t \to \infty$. This corresponds to a situation where a proportion ϵ_j of the channels at node j are considered "saturated", i.e. their queues are infinitely long and remain so always, irrespective of arrivals or departures. Of course $\underline{\mathbf{a}}^{(1)}$, given by $a_j^{(1)}(n) = 1\ \forall j,n$, is also a fixed point; here every channel in the network is saturated.

4 Convergence of Stationary Distributions

We first prove Theorem 3, from which we can deduce the existence and various properties of the stationary distributions of the Markov processes $\underline{\mathbf{r}}_N(t)$ under the non-overload condition (1.7).

Proof of Theorem 3: The argument is an extension of that used by Turner to prove Theorem 4 in [7]. For a network S, and $r \geq 1$, $1 \leq k \leq J$, let $A_{k,r}(S)$ be the time of the rth arrival after time 0 (from outside or from within the network) at node k, and $D_{k,r}(S)$ the time of the rth departure from node k.

We will construct a coupling such that for all k, r

$$\begin{aligned} &\text{(i)} \quad D_{k,r}(R) \leq D_{k,r}(Q), \\ &\text{(ii)} \quad A_{k,r}(R) \leq A_{k,r}(Q). \end{aligned} \tag{4.1}$$

As well as the real tasks in Q, we invent "shadow" tasks, whose behaviour links the networks Q and R. For each k and r, a shadow task arrives at node k in Q at time $A_{k,r}(R)$, and remains until time $A_{k,r}(Q)$, when it is replaced by the arriving real task. This definition makes sense if (4.1)(ii) holds. The shadow task inhabits the same channel at node k in Q as the real task which will replace it. We will write $\overline{Q}$ to mean the network Q including shadows as well as real tasks; thus a node in the network $\overline{Q}$ experiences arrivals at the same time as the corresponding node in R, and departures at the same time as that in Q.

For a network S, and $x \in \mathbb{Z}_+$, define

$$\psi_{k,x}(S,t) = \sum_{n=1}^{N} [l_{k,n}(S,t) - x]_+,$$

where $l_{k,n}(S,t)$ is the queue length of the nth channel at node k in S at time t, and $[y]_+ = \max(y, 0)$.

In order to show (4.1), we will also establish that, under the coupling,

$$\psi_{k,x}(R,t) \leq \psi_{k,x}(\overline{Q},t) \tag{4.2}$$

for all $x \in \mathbb{Z}_+$, $t \geq 0$, $1 \leq k \leq J$.

The following observation from [7] will be used here. The difference between $\psi_{k,x}(S,t)$ and $\psi_{k,x+1}(S,t)$ is equal to the number of queues at node k in S with length greater than x at time t. Thus if $\psi_{k,y}(R,t) \leq \psi_{k,y}(\overline{Q},t)$ for all y, and $\psi_{k,x}(R,t) = \psi_{k,x}(\overline{Q},t)$, then it follows from the inequality (4.2) at $x+1$ and $x-1$ that

$$\#\{n : l_{k,n}(R,t) \leq x\} \leq \#\{n : l_{k,n}(\overline{Q},t) \leq x\} \tag{4.3}$$

and

$$\#\{n : l_{k,n}(R,t) \geq x\} \leq \#\{n : l_{k,n}(\overline{Q},t) \geq x\}. \tag{4.4}$$

The coupling works as follows. For each k, generate a Poisson process of rate $N\lambda_k$ for the external arrival times at node k, and one of rate $N\mu_k$ for the potential departure times from node k. These are the same in both systems. Also, the rth task to depart from node k in Q will have the same destination as the rth task to depart from node k in R (generated i.i.d. according to the probabilities $p_{k,j}$ and p_k^*). This ensures that if (4.1)(i) holds for the departure process up to a given time, then so does (4.1)(ii) for the arrival process up to that time. We now need to describe the way in which channels within nodes are chosen for arrivals and departures, verifying that departures preserve (4.1) and that all network events preserve (4.2). (Certainly (4.2) holds at $t = 0$ since the initial states of $\overline{Q}$ and R are the same).

At a potential departure time at node k, rank the channels at k in $\overline{Q}$ and in R in order of queue lengths (including shadow tasks), and let a departure occur from the correspondingly ranked queue in each system; for example, if it occurs from the longest queue in $\overline{Q}$, let it occur from the longest queue in R also. (Departures from an empty queue, or from one containing only shadow tasks, are lost). Now let D be a potential departure time at node k, and suppose that properties (4.1) and (4.2) hold for $t < D$; we wish to show that they are preserved. Firstly, the only way in which (4.2) could be violated by a departure point at node k would be if $\psi_{k,x}(R, D-) = \psi_{k,x}(\overline{Q}, D-)$ for some x, and if queues of length a in $\overline{Q}$ and length b in R were chosen for departure at time D, with $0 \leq b \leq x < a$. But this is impossible, since (4.3) holds for $t < D$ and since the ranks of the channels chosen for departure are coupled. Secondly, for (4.1)(i) to be violated, the number of previous departures from node k before time D must be the same in R as in Q (and hence also $\overline{Q}$). Then, since arrivals in R and $\overline{Q}$ occur at the same times, there must be the same total number of tasks at node k in R and in $\overline{Q}$; i. e. $\psi_{k,0}(R, D-) = \psi_{k,0}(\overline{Q}, D-)$. But then, by (4.3) again, there are at least as many empty queues in $\overline{Q}$ as in R. Thus if a departure occurs from a non-empty queue in $\overline{Q}$, the coupling of the ranks of the channels chosen for departure ensures that a departure must also occur from a non-empty channel of R, preserving (4.1)(i) and hence also (4.1)(ii).

At an arrival time $A = A_{k,r}(R)$, generated either by an external arrival or by a departure within the network, rank the channels at node k in $\overline{Q}$ and in R by queue length as before. Choose $m(R)$ channels at random and send the arrival in R to the one with the shortest queue. Choose one of the correspondingly ranked $m(R)$ channels in $\overline{Q}$ at random and send the shadow task to this channel. The effect is that if an arrival occurs to the nth longest queue in R, the corresponding arrival in $\overline{Q}$

must occur to a queue at least as the long as the nth longest there. Again, suppose (4.2) has held up to this point - how might it now fail? If the chosen queue in $\overline{Q}$ has length a, and that in R length b, then it would require $\psi_{k,x}(R, A-) = \psi_{k,x}(\overline{Q}, A-)$ to hold for some x, with $a < x \leq b$. Then, following (4.4), we would have

$$\#\{n : l_{k,n}(R, A-) \geq x\} \leq \#\{n : l_{k,n}(\overline{Q}, A-) \geq x\};$$

but this is impossible, since $a < x \leq b$ and, according to our coupling scheme, we must have

$$\#\{n : l_{k,n}(R, A-) \geq b\} \geq \#\{n : l_{k,n}(\overline{Q}, A-) \geq a\},$$

(since a queue with length a in $\overline{Q}$ is at least as long as one whose ranking corresponds to a queue with length b in R), and $\#\{n : l_{k,n}(\overline{Q}, A-) = a\} \geq 1$ (a queue of length a must exist in $\overline{Q}$ since a queue of length a has been chosen there for the arrival).

Finally, at an arrival time $A_{k,r}(Q)$, the shadow task which arrived at time $A_{k,r}(R) \leq A_{k,r}(Q)$ is replaced by a real task in the same channel. This does not affect the queue lengths in $\overline{Q}$ so (4.2) is preserved.

Note that in fact no time elapses between a departure and an arrival which it generates at the same or another node; we have separated the two events in the above description, but if the two separately do not violate the properties (4.1) and (4.2) then nor will the two combined. Similarly, it may be that $A_{k,r}(R) = A_{k,r}(Q)$ for some k and r; in this case the shadow task arriving at this time is never seen since it is instantaneously erased by the arriving real task.

Under the coupling described, arrivals in the network from outside occur at the same time in R as in Q, and departures from the network to the outside occur no later in R than in Q; hence the total number of tasks in R at a given time is no greater than that in Q. $\square$

Remark: We have not resolved the question of whether the conclusion of Theorem 3, or a similar one, holds also in the case $1 < m(Q) < m(R)$. The complication introduced by the network structure is that if one part of the network works more "slowly", this can lead to delayed arrivals in another part; and these delayed arrivals may, in the light of more recent information, be able to make a better choice about which queue to join than if they had arrived earlier.

Proof of Theorem 4:

(i) If $m = 1$, then the network described by $\underline{\mathbf{r}}_N(t)$ is equivalent to a standard Jackson network, with NJ nodes $(j, n), 1 \leq j \leq J, 1 \leq n \leq N$, arrival and service rates $\lambda_{j,n} = \lambda_j$ and $\mu_{j,n} = \mu_j$, and routing matrix $p_{(j,n)(j',n')} = p_{jj'}/N$. Then $\rho_{(j,n)} = \rho_j$ also, and under the condition (1.7) the network is positive recurrent, with an invariant distribution analogous to (1.8).

For $m > 1$, Theorem 3 allows us to couple the network with a "slower" one for which $m = 1$ in such a way that the faster network is empty whenever the slower one is empty. Then, since the slower network is positive recurrent, the faster one must be also.

(ii) Since $\overline{\mathcal{U}}^J$ is compact, so is the set $\mathcal{P}(\overline{\mathcal{U}}^J)$ of probability measures on $\overline{\mathcal{U}}^J$. Hence the sequence of probability measures $\{\pi_N, N \in \mathbb{N}\}$ has limit points. We wish to check that any limit point is the Dirac delta-measure concentrated at $\underline{\mathbf{a}}$.

From Lemma 3.2, the maps $\underline{\mathbf{g}} \mapsto \underline{\mathbf{u}}(t, \underline{\mathbf{g}}), t \geq 0, \underline{\mathbf{g}} \in \overline{\mathcal{U}}^J$ are continuous in $\underline{\mathbf{g}}$, and it follows from Theorem 2 that any limit point π of $\{\pi_N\}$ must be an invariant distribution for these maps. Hence, by Theorem 1(iii), it is sufficient to show that π is concentrated on $\mathcal{U}^J$, i.e. $\pi(\{\underline{\mathbf{g}} : v_j(1, \underline{\mathbf{g}}) < \infty \ \forall j\}) = 1$. We show in fact that $\mathbb{E}_\pi v_j(1) < \infty$ for all j. If $m = 1$, $\mathbb{E}_{\pi_N} r_j(n) = (\rho_j/\mu_j)^n$ for all N, by the argument in the proof of part (i) and the expression (1.8), so that $\mathbb{E}_{\pi_N} v_j(1) = \rho_j/(\mu_j - \rho_j)$ for all N. Hence, by Theorem 3 and standard results on the convergence of Markov processes to equilibrium, $\mathbb{E}_{\pi_N} v_j(1) \leq \rho_j/(\mu_j - \rho_j)$ for all N and any m, giving $\mathbb{E}_\pi v_j(1) \leq \rho_j/(\mu_j - \rho_j)$ also, as required. $\square$

Acknowledgements. We would like to thank the referee for careful reading and helpful remarks, and Ed Thomas for some valuable comments. YMS would like to thank Prof. F. I. Karpelevich, Dr. N. D. Vvedenskaya and Dr. A. N. Rybko for encouraging and critical discussions, and also I.H.E.S., Bures-sur-Yvette, France, for their warm hospitality during visits in 1997.

References

[1] Ethier, S. N. and Kurtz, T. G., (1985) *Markov Processes: Characterization and Convergence.* John Wiley & Sons.

[2] Jackson, J. R., (1957) Networks of Waiting Times. *Operational Research* **5**, 518–527.

[3] Jackson, J. R., (1965) Jobshop-like Queueing Systems. *Management Sci.* **10**, 131–142.

[4] Karpelevich, F. I., Pechersky, E. A. and Suhov, Y. M., (1996) Dobrushin's Approach to Queueing Network Theory. *J. Appl. Math. Stoch. Anal.* **9**, 373–397.

[5] Kelly, F. P., (1991) Loss Networks. *Annals of Applied Probability* **1**, 319–378.

[6] Mitzenmacher, M., (1996) The Power of Two Choices in Randomized Load Balancing. Ph.D. thesis, University of California, Berkeley.

[7] Turner, S. R. E., (1998) The Effect of Increasing Routing Choice on Resource Pooling. *Probability in the Engineering and Informational Sciences* **12**, 109–124.

[8] Vvedenskaya, N. D., Dobrushin, R. L. and Karpelevich, F. I., (1996) Queueing System with Selection of the Shortest of Two Queues: An Asymptotic Approach. *Problems of Information Transmission* **32**, 15–27.

[9] Vvedenskaya, N. D. and Suhov, Y. M., (1997) Dobrushin's Mean-Field Approximation for a Queue with Dynamic Routing. *Markov Processes and Related Fields* **3**, 493–526.

Département de Mathématiques
et d'Informatique,
Ecole Normale Supérieure,
45 rue d'Ulm,
75230 PARIS Cédex 05,
FRANCE
James.Martin@ens.fr

Statistical Laboratory,
DPMMS,
University of Cambridge,
16 Mill Lane,
Cambridge CB2 1SB,
UK
yms@statslab.cam.ac.uk