The Annals of Applied Probability
1994, Vol. 4, No. 1, 124–148

STABILITY OF GENERALIZED JACKSON NETWORKS*

By S. P. Meyn and D. Down

University of Illinois at Urbana-Champaign

In this paper we study open generalized Jackson networks with general arrival streams and general service time distributions. Assuming that the arrival rate does not exceed the network capacity and that the service times possess conditionally bounded second moments, we deduce stability of the network by bounding the expected waiting time for a customer entering the network. For Markovian networks we obtain convergence of the total work in the system, as well as the mean queue size and mean customer delay, to a unique finite steady state value.

1. Introduction. The study of (generalized) Jackson networks can be traced back to their namesake [13], who considered networks with Poisson inputs and exponential service times and showed that the invariant probability for the process has a simple product form. Several generalizations of these results are derived by Kelly [15]. The foregoing assumptions on the arrival streams and service times were made to greatly simplify the analysis of these networks. The relaxing of these assumptions was the subject of the work by Borovkov [4], where a model similar to our Markovian network is considered. The finite buffer case is treated in Konstantopoulos and Walrand [18], and general point process arrival streams and general service processes are considered for networks without feedback [17]. Ergodicity for closed generalized Jackson networks is treated by Kaspi and Mandelbaum [14].

Cruz [5, 6] has performed an analysis of deterministic multiclass networks, but in the analysis there is difficulty handling feedback of customers within the network, a situation that we will address. Also in a deterministic setting are the results of Kumar and co-workers [20, 22, 30] that consider the stability properties of general multiclass networks.

Sigman considers open queueing networks [32]. This work has brought forth several ideas that have proved important in the methods of analysis contained in this paper.

In this paper we devise a stability proof based upon induction on the number of nodes in the network. This simplifies the analysis of multinode networks, and the test function approach used eliminates the need to search for the existence of a regeneration time [2] or a suitable "small subset" of the state space [27, 29]. From this proof we obtain mean boundedness of the queue lengths and under slightly stronger conditions we show that the

Received November 1991; revised May 1993.

[1]Supported in part by University of Illinois Research Board Grant 1-2-69637.

AMS 1991 *subject classifications*. 68M20, 60J10.

Key words and phrases. Queueing networks, Harris recurrence, general state space Markov processes.

expectation of the queue lengths, customer delay and total work converge to finite steady state values. Some of the results reported here were previously published in abridged form [24]. The general form of the test function that we use is suggested in Kingman [16], where quadratic functions are used to characterize ergodicity of random walks on an orthant. Quadratic functions have recently been used to obtain performance bounds [31] and conditions for stability [19] for general exponential queueing networks.

The organization of the paper is as follows. The first part of Section 2 contains a complete analysis of the single queue network. This is the basis for the rest of the section, in which an inductive argument is used for a network of queues. We show that the delay expected by a customer entering the network is bounded, the results being obtained for non-Markovian networks for which the interarrival times are not necessarily independent and identically distributed (i.i.d.). Section 3 considers arrival streams that are renewal processes, in which case we model the system as a continuous time parameter Markov process. Recent work by Meyn and Tweedie [25, 26, 28] is used to obtain Harris ergodicity for the network, as well as mean boundedness of WIP (work in progress) and queue size.

2. Stability of open Jackson networks.

2.1. *The single queue.* We begin with the simplest case, a single queue. In this situation, we have customers arriving at the system with local rate γ and, after waiting for service, they are processed by a server with service rate μ. Upon completing service, a customer leaves the system with probability $r_0 > 0$, and reenters the queue with probability $1 - r_0$.

The load condition that must be satisfied in this case is clearly

$$\rho := \frac{\lambda}{\mu} < 1, \tag{1}$$

where $\lambda = \gamma / r_0$.

Let $\Phi_s := \binom{Q_s}{B_s}$, $s \in \mathbb{R}_+$, where, at time s, Q_s denotes the queue length, not including the customer in service, and B_s is the time the current customer has been in service. We let N_t denote the number of exogenous arrivals that occur in the interval $(0, t]$, and $\mathscr{F}_s$ denote a given filtration that is richer than $\sigma\{N_r, \Phi_r : r \le s\}$ and for which all of the service times and routing initiated at a time greater than s are independent of $\mathscr{F}_s$.

The following standing assumptions will be made concerning the arrival stream and the system.

ASSUMPTION A1. There exist deterministic constants γ and L such that

$$\mathsf{E}[N_{t+s} - N_s | \mathscr{F}_s] \le \gamma t + L, \qquad s, t \in \mathbb{R}_+.$$

ASSUMPTION A2. The arrival stream, service times and customer routing are mutually independent. The service times are i.i.d. with common mean $1/\mu$ and finite second moment.

Assumption A1 describes precisely how the local rate of customers to the system is bounded by γ. For a renewal process with mean interarrival time γ^{-1}, it follows from the strong Markov property and ergodicity of the forward process that A1 is satisfied (cf. [2]). Previous papers [5, 6, 22] assume a deterministic version of A1 without a conditional expectation.

It is important to note here that A1 is called the local rate of arrival because it rules out the phenomenon of a "large" number of arrivals occurring in small intervals of time. This is as opposed to infinite time horizon or long-term formulations of rate such as

$$\limsup_{t\to\infty} \frac{N(t)}{t}. \tag{2}$$

An example will serve to illustrate the consequences of considering A1 as a definition of arrival rate.

Consider a single server with service times deterministic of length 2. The arrival sequence is described by a regenerative sequence of interarrival times $T(n)$, where a cycle is defined as follows: Let N be a nonnegative integer-valued random variable with finite first moment and infinite second moment. There are $N+1$ arrivals during a cycle. Now, for $j = 1,\dots,N$, $T(j) = 1$ and $T(N+1) = N+3$. The work load is a positive recurrent regenerative process, because the long-term arrival rate as defined by (2) is less than the service rate. This is a result of the last interarrival time within the cycle being so large as to empty the system. However, the average delay in the system is infinite, as demonstrated in the following calculation. For a regenerative cycle, define $D(n)$ as the delay in queue of arrival n. It is easy to see that for $j = 1,\dots,N$, $D(j) = j$ and $D(N+1) = 0$. Thus the average delay is given by $\mathsf{E}[N(N+1)/2]/\mathsf{E}[N]$, which is infinite. This arrival process satisfies A1, but with $\gamma = 1$, which of course is not small enough for our purposes.

Assumption A2 may be generalized somewhat; for instance, the condition that the service times are i.i.d. can be substantially relaxed. We do not consider this more general framework in this paper.

Our goal is to bound the mean waiting time for a customer entering service, which is closely related to the total work in the system, measured in units of time. This objective forces us to assume that the service times possess bounded second moments. If this condition is violated, then the expected queue size in steady state may be infinite, so we cannot expect to obtain uniform bounds on the expected work in the system in this case (cf. [2], Theorem 2.1, page 184).

We denote the total work in the system at time t by W_t, which is the total amount of service time that the customers that are in the queue at time t will receive. We note that $\mathsf{E}[W_t]$ is equal to the expected amount of time required for the system to empty from time t, given that no new arrivals occur.

The main result of this section shows that our conditions imply that the network is stable in the mean. In the next subsection, this result is generalized to multidimensional networks by induction on the number of nodes in the system.

THEOREM 2.1. *If Assumptions A1 and A2 hold, and if the load condition (1) is satisfied, then there exist constants $T, b < \infty$, that are independent of the initial condition of the network, such that*

$$\frac{1}{R}\int_0^R \mathsf{E}[W_s]\,ds \le b + \frac{1}{R}\int_0^T \mathsf{E}[W_s^2]\,ds, \qquad R \in \mathbb{R}_+.$$

In particular,

$$\limsup_{R\to\infty} \frac{1}{R}\int_0^R \mathsf{E}[W_s]\,ds \le b.$$

The proof is postponed to the end of this subsection.

Over the time interval $(s, t]$ we define the total change in work $W_t - W_s$ and the total work to enter the system $\Omega_{s,t}$ by

$$(3) \qquad W_t - W_s = \Omega_{s,t} - T_{s,t}, \qquad \Omega_{s,t} = \sum_{N_s < j \le N_t} \sum_{k=1}^{R_j} S_{j,k},$$

where R_j is the number of times that customer j revisits the system, $T_{s,t}$ is the amount of time that the system is busy in $[s, t]$ and the service times $\{S_{j,k}\}$ are i.i.d. with common mean $1/\mu$.

We may arrive at a bound on the work arriving at the system, similar to our assumption on the arrival stream, using Assumptions A1 and A2:

$$(4) \qquad \mathsf{E}[\Omega_{s,t}|\mathscr{F}_s] \le \frac{\lambda}{\mu}(t-s) + \bar{L} = \rho(t-s) + \bar{L},$$

where $\bar{L} = L/\mu$.

We may now state and prove the following negative drift property for the expected work in the queue:

LEMMA 2.2. *There exist constants T and b, independent of $s \in \mathbb{R}_+$, such that for all initial buffer levels Q_0 and residual service times Y_0,*

$$\mathsf{E}\left[\frac{1}{T}\int_0^T W_{s+t}^2\,dt \,\Big|\, \mathscr{F}_s\right] \le \mathsf{E}[W_s^2|\mathscr{F}_s] - \mathsf{E}[W_s|\mathscr{F}_s] + b.$$

PROOF. From (3) we have

$$
\begin{aligned}
\mathsf{E}\big[W_{s+t}^2|\mathscr{F}_s\big] &= \mathsf{E}\big[(W_s + \Omega_{s,s+t} - T_{s,s+t})^2|\mathscr{F}_s\big] \\
&= \mathsf{E}\big[W_s^2|\mathscr{F}_s\big] + \mathsf{E}\big[(\Omega_{s,s+t})^2|\mathscr{F}_s\big] + \mathsf{E}\big[T_{s,s+t}^2|\mathscr{F}_s\big] \\
&\quad + 2\mathsf{E}\big[W_s\Omega_{s,s+t}|\mathscr{F}_s\big] - 2\mathsf{E}\big[W_sT_{s,s+t}|\mathscr{F}_s\big] \\
&\quad - 2\mathsf{E}\big[\Omega_{s,s+t}T_{s,s+t}|\mathscr{F}_s\big] \\
&\le \mathsf{E}\big[W_s^2|\mathscr{F}_s\big] + 2\mathsf{E}\big[W_s|\mathscr{F}_s\big](\rho t + \bar{L}) \\
&\quad - 2\mathsf{E}\big[W_sT_{s,s+t}|\mathscr{F}_s\big] + O(t^2),
\end{aligned}
\tag{5}
$$

where the term $O(t^2)$ is deterministic and does not depend on s.

Ignoring the work that arrives from outside the network in $(s, s+t]$ gives the bound $\mathsf{E}[W_sT_{s,s+t}|\mathscr{F}_s] \ge \mathsf{E}[W_s(W_s \wedge t)|\mathscr{F}_s] \ge t\mathsf{E}[W_s|\mathscr{F}_s] - t^2$, where the second inequality follows from applying the bound

$$x(x \wedge T) \ge xT - T^2,$$

valid for any $x, T > 0$. The inequality (5) then gives

$$
\begin{aligned}
\mathsf{E}\big[W_{s+t}^2|\mathscr{F}_s\big] &\le \mathsf{E}\big[W_s^2|\mathscr{F}_s\big] + 2(\rho t + \bar{L})\mathsf{E}\big[W_s|\mathscr{F}_s\big] \\
&\quad - 2t\mathsf{E}\big[W_s|\mathscr{F}_s\big] + O(t^2).
\end{aligned}
\tag{6}
$$

Integrating both sides over t we obtain the result. □

For stability we apply the following result:

LEMMA 2.3. *Let a_s, b_s, $s \in \mathbb{R}_+$, be positive finite-valued measurable functions on $\mathbb{R}_+$. If for some fixed $B > 0$, $T > 0$,*

$$\frac{1}{T}\int_0^T a_{s+t}\,dt \le a_s - b_s + B, \qquad s \in \mathbb{R}_+, \tag{7}$$

then for all $R > 0$,

$$\frac{1}{R}\int_0^R b_t\,dt \le B + \frac{1}{R}\int_0^T a_t\,dt.$$

PROOF. Let $R > 0$ and average (7) over s to obtain

$$
\begin{aligned}
\frac{1}{R}\int_0^R a_s\,ds - \frac{1}{R}\int_0^R b_s\,ds + B &\ge \frac{1}{RT}\int_0^T\left\{\int_0^R a_{s+t}\,ds\right\}dt \\
&= \frac{1}{RT}\int_0^T\left\{\int_0^R a_s\,ds - \int_0^t a_s\,ds + \int_R^{R+t} a_s\,ds\right\}dt \\
&\ge \frac{1}{R}\int_0^R a_s\,ds - \frac{1}{R}\int_0^T a_s\,ds.
\end{aligned}
$$

Cancelling and rearranging terms gives the result. □

We conclude this section with the proof of Theorem 2.1.

PROOF OF THEOREM 2.1. The result follows immediately from Lemmas 2.2 and 2.3 on taking $a_s = \mathsf{E}[W_s^2]$, $b_s = \mathsf{E}[W_s]$. □

We remark that in the single queue case, the average over T is not necessary. Equation (6) is stronger than Foster's criterion and may be shown to imply Theorem 2.1 directly (see [21, 27]). The reason for taking this step is that, since this step is necessary in our stability proof for networks, it is convenient to illustrate the technique first on this simple model.

2.2. *Networks of queues.* We now consider networks composed of a collection of d queues. Customers arrive in d separate arrival streams, one for each node in the network. A customer waits in its node of arrival until it can obtain service, and is then routed to another buffer in the network under a random routing policy. This procedure repeats itself until the customer leaves the network.

The required service time at queue $i \in I := \{1, \dots, d\}$ has a general distribution with mean $1/\mu_i$, and customers arrive to node i from outside the network at local rate γ_i.

Whenever a job has completed service at queue i, it is routed to queue j with probability r_{ij} and leaves the system with probability r_{i0}. The network is assumed *open*, that is, every job entering the system may leave the system with positive probability and will be called a (*generalized*) *Jackson network*.

Let Q_t denote the vector whose components indicate the queue lengths (not including the customers that may be in service) of the buffers in the system and let the vector B_t denote the times that the current customers have been in service (set to zero if the server is free) at time t. The stochastic process $\Phi_t = \binom{Q_t}{B_t}$, $t \in \mathbb{R}_+$, is thus seen to evolve on $\mathsf{X} := \mathbb{Z}_+^d \times \mathbb{R}_+^d$.

Let $N_t(i)$ denote the number of exogenous arrivals that occur in the interval $(0, t]$ at node i and let $Y_t(i)$ be the residual service time at node i (set to zero if the server is free). Let $(\mathscr{F}_r)$ be a filtration for which the stochastic process $\{\Phi_r\}$ is $(\mathscr{F}_r)$-adapted and for which all of the service times and routing initiated at a time greater than r are independent of $\mathscr{F}_r$. We also require $\mathsf{E}[Y_s(i)|\mathscr{F}_s] = \mathsf{E}[Y_s(i)|B_s(i)]$.

Our conditions on the network are made precise in the following assumptions:

ASSUMPTION B1. For each $i \in I$, there exist deterministic constants γ_i and L_i such that

$$\mathsf{E}\big[N_{t+s}(i) - N_s(i)\,|\mathscr{F}_s\big] \le \gamma_i t + L_i, \qquad s, t \in \mathbb{R}_+.$$

ASSUMPTION B2. The arrival streams, service times and customer routing are mutually independent. The service times at each node are i.i.d. with common mean $1/\mu_i$.

ASSUMPTION B3. The network is open: Any customer entering the network may leave it.

ASSUMPTION B4. At each node $i \in I$ we have for a generic service time $S_1(i)$,

$$\mathsf{E}\left[S_1(i) - t \mid S_1(i) > t\right] \le \overline{S}, \qquad t \ge 0,$$

where $\overline{S}$ is a deterministic constant.

As noted before, Assumption B1 is easily verified if we specialize our arrival processes to be renewal processes.

Through condition B4, we have imposed stronger conditions on the service times. This is to avoid a situation known as blocking. In our proof, we will consider the case when the state of the system is "large" and prove a drift condition for the network. We will see that if at one node the residual service time is large, then work at that node will decrease but the rest of the network may be starved for work.

It may be shown that Assumption B4 is equivalent to condition (1) of Borovkov [4]. Indeed, the calculation on page 414 of [4] implies that under B4, or condition (1) of [4], there exist constants $\overline{E}$, δ, independent of $t \in \mathbb{R}_+$, such that

$$\mathsf{E}\left[\exp(\delta(S_1(i) - t)) \mid S_1(i) > t\right] \le \overline{E}.$$

This bound together with the assumptions on the σ-algebra $\mathscr{F}_s$ imply that we also have the following bound for the residual service times:

$$\begin{aligned}\mathsf{E}\left[\exp(\delta Y_s(i)) \mid \mathscr{F}_s\right] &= \mathsf{E}\left[\exp(\delta(S_1(i) - t)) \mid S_1(i) > t\right] \circ (t = B_s(i)) \\ &\le \overline{E}, \qquad i \in I,\ s \ge 0.\end{aligned}$$

Let $\{\lambda_i\}$ denote the constants defined by the traffic equations

$$\lambda_i := \gamma_i + \sum_{j \in I} \lambda_j r_{ji}.$$

When the system is stable, the local rate at which customers arrive at the ith queue is equal to λ_i. The load condition that we will now assume to be satisfied is

$$\rho_i := \frac{\lambda_i}{\mu_i} < 1 \quad \text{for each } i \in I. \tag{8}$$

In a similar manner as for the single queue, $W_t(i)$ is defined as the total service time that node i will use to service all of the customers that are in the *system* at time t. We stress that this quantity includes the time that node i will spend servicing a customer that is at a different node at time t, but will eventually arrive at node i. The total work in the system at time t is given by

$$W_t := \sum_{i \in I} W_t(i).$$

The main result of this section is the following generalization of Theorem 2.1. This follows directly from the drift property obtained in Lemma 2.5.

THEOREM 2.4. *If Assumptions* B1–B4 *hold and if the load condition* (8) *is satisfied, then there exist constants* $T, b < \infty$, *that are independent of the initial condition of the network, such that*

$$\frac{1}{R}\int_0^R \mathsf{E}[W_s]\,ds \le b + \frac{1}{R}\int_0^T \mathsf{E}[W_s^2]\,ds, \qquad R \in \mathbb{R}_+.$$

In particular,

$$\limsup_{R\to\infty} \frac{1}{R}\int_0^R \mathsf{E}[W_s]\,ds \le b.$$

In Section 3 we will show that when the input to the system is a spread-out renewal process, the expectation $\mathsf{E}[W_s]$ converges to a finite steady state value. Hence, in this case the average over $s \in \mathbb{R}_+$ is not necessary.

The following functions on X will play the role of Lyapunov functions in our stability proof:

$$\begin{aligned} V_2(\Phi_t) &= \sum_{i=1}^{d} \mathsf{E}\big[W_t(i)\,|\mathscr{F}_t\big]^2, \\ V_1(\Phi_t) &= \sum_{i=1}^{d} \mathsf{E}\big[W_t(i)\,|\mathscr{F}_t\big]. \end{aligned} \tag{9}$$

Because future service times and customer routing are assumed independent of $\mathscr{F}_t$, it is easy to see, by Assumption B2, that V_1 and V_2 are well defined functions on X.

Let $\Omega^l_{s,t}(i)$ denote the work arriving from arrival stream l, in the time interval $(s,t]$, to be completed by node i. This quantity may be written explicitly as

$$\Omega^l_{s,t}(i) = \sum_{N_s(l) < j \le N_t(l)} \; \sum_{k=1}^{R^l_j(i)} S^l_{j,k}(i),$$

where $N_s(l)$ is the number of arrivals from the lth stream in $(0,s]$, the service times $\{S^l_{j,k}(i)\}$ at node i are i.i.d. random variables with mean $1/\mu_i$ and $R^l_j(i)$ denotes the number of times that the jth customer from arrival stream l visits node i.

We let $\Omega_{s,t}(i)$ denote the total work that is destined for node i, arriving in the time interval $(s,t]$:

$$\Omega_{s,t}(i) = \sum_{k=1}^{d} \Omega^k_{s,t}(i).$$

We may write a bound similar to (4) for the total expected work arriving to the system in the time interval $(s,t]$, to be completed by node i:

$$\mathsf{E}\left[\Omega_{s,t}(i)\,|\,\mathscr{F}_s\right] \le \rho_i(t-s) + \bar{L}_i, \tag{10}$$

where $\bar{L}_i = L_i/\mu_i$. This bound follows from our assumptions on $(\mathscr{F}_s)$, Assumptions B1 and B2 and the traffic equations.

We find that centering the random variable $V_2(\Phi_{s+t})$ simplifies the analysis, allowing us to examine a simpler term than $V_2(\Phi_{s+t})$ itself. The centering is as follows:

$$\begin{aligned} &\mathsf{E}\left[V_2(\Phi_{s+t})\,|\,\mathscr{F}_s\right] \\ &\quad = \mathsf{E}\left[\left(\mathsf{E}[W_{s+t}|\mathscr{F}_{s+t}] - \mathsf{E}[W_{s+t}|\mathscr{F}_s] + \mathsf{E}[W_{s+t}|\mathscr{F}_s]\right)^2\Big|\mathscr{F}_s\right] \\ &\quad = \mathsf{E}\left[\left(\mathsf{E}[W_{s+t}|\mathscr{F}_{s+t}] - \mathsf{E}[W_{s+t}|\mathscr{F}_s]\right)^2\Big|\mathscr{F}_s\right] + \mathsf{E}[W_{s+t}|\mathscr{F}_s]^2. \end{aligned} \tag{11}$$

By Assumptions B1 and B2, the first term on the right-hand side of (11) is $O(t^2)$, so we need analyze only the second term, for which it is easy to obtain a drift.

The following result together with Lemma 2.3 and Jensen's inequality immediately gives Theorem 2.4.

LEMMA 2.5. *Suppose that conditions* B1–B4 *and the load condition* (8) *are satisfied. Then there exist constants T and b, which depend only on the service time distributions, the constants (λ_i) and (L_i), the routing and topology of the network, and, in particular, are independent of $s \in \mathbb{R}_+$, such that for all initial buffer levels and residual service times,*

$$\mathsf{E}\left[\frac{1}{T}\int_0^T V_2(\Phi_{s+t})\,dt\,|\mathscr{F}_s\right] \le V_2(\Phi_s) - V_1(\Phi_s) + b \quad a.s. \tag{12}$$

An inductive proof of this result comprises the remainder of this section.

Applying a centering argument, as in (11), and Jensen's inequality to the bound obtained in Lemma 2.2 we see that (12) holds in the single node case. For the general d-node case we will isolate a single node, which we will take as node 1, whose buffer is essentially "full." For this node, a drift property of the form (12) has essentially been established in Lemma 2.2. For the remainder of the network, we view the output of node 1 as an exogenous input. Up until the first time that node 1 is empty, this fictitious input supplies an adequate amount of work to the remainder of the network. By induction, the $(d-1)$ node network processes work at the rate it arrives. We will show that this reasoning leads to (12).

The corollary to Proposition 2.6 justifies the notion that queue 1 supplies an adequate amount of work to the remainder of the network, and Lemma 2.7 shows that a stable network completes work at the rate it arrives.

Consider a duplicate of the network under study here with state process Φ'_t and busy time $T'_{0,t}$, $t \in \mathbb{R}_+$. We assume that the two networks are identical, except that $Q_0(j) \geq Q'_0(j)$, $j \in I$. Specifically, in each network, the jth customer to arrive at node i will require the same amount of service and upon completion of service, will be routed to the same machine.

PROPOSITION 2.6. *The busy times for the two networks are related by* $T_{0,t}(i) \geq T'_{0,t}(i)$ *a.s.*

PROOF. We first note that the problem is a deterministic one, as we fix common sample paths for the service times, arrival processes and customer routing.

Let t_k denote the kth instant at which a service is completed in either the new or the original network. The variables $t_j(i)$ and $t'_j(i)$ are the times of the jth service completion at node i for the original and new networks, respectively. We are also concerned with the number of arrivals to a specific node by time t, which we define to be the total number of customers that are in queue or in service at time t, plus those that have already been serviced by time t. The proof will be by induction, with the following induction hypothesis:

At time t_k, for all $i \in I$:

(a) For each j, $t_j(i) \leq t'_j(i)$ when $t_j(i), t'_j(i) \leq t_k$.

(b) {Number of arrivals to node i in $[0, t_k]$ for the original network} $\geq$ {number of arrivals to node i in $[0, t_k]$ for the new network}.

Once we perform the proof of this hypothesis, the proposition follows easily.

Consider $k = 1$. The result (a) is true because the first service completion will be due to either a customer with positive residual service time or a customer arriving from outside the system to an empty queue. The result (b) follows from (a), the fact that the outside arrival streams are identical for both networks, and the assumption on the initial queue sizes.

We will now assume the result for $1 \leq j \leq k$ and show that it is true for $k + 1$. There are two cases to be examined here. Fix attention to a single node $j \in I$.

If a service is completed at node j at time t_{k+1} in the original network, the result (a) is clear. For the result (b), we see that it holds from (a) that the outside arrival streams are identical for both networks and that we have fixed a common sample path for customer routing in both networks.

If a service is completed at t_{k+1} for the lth customer serviced at node j in the new network, then we have (a), as if we look back to when the lth customer arrived to the queue, we can conclude that as outside arrivals occur at identical times and as (by the induction hypothesis) arrivals from within the network occur earlier in the original network, the lth arrival occurs earlier in the original network, implying (a). As before, the result (b) is then obvious. □

Proposition 2.6 also yields the following related result. Analogous to the construction for Proposition 2.6, we consider a duplicate of the network under study with state process Φ^*_t and busy time $T^*_{0,t}$, $t \in \mathbb{R}_+$. We assume that the two networks are identical, except that $N_t^*(j) = N_t(j) + \Delta N_t(j)$ and $Q_0^*(j) = Q_0(j) + \Delta Q_0(j)$, $1 \le j \le d$, where $\Delta Q_0(j) \ge 0$ and $\Delta N_t(j)$ is increasing, $j \in I$.

COROLLARY. *The busy times for the two networks are related by* $T^*_{0,t}(i) \ge T_{0,t}(i)$, $i \in I$, $t \ge 0$.

PROOF. Create a (fictitious) queue with buffer levels $Q_t(d+1)$ and $Q_t^*(d+1)$ in the original and new networks, respectively. Assume that $Q_0^*(d+1) = \infty$ and that service times and routing from node $d+1$ are such that parts are released and routed to account for the differences in arrival streams [i.e., n parts are released from the fictitious queue and routed to queue j every time $\Delta N_{t^+}(j) - \Delta N_t(j) = n$]. If we also assume that $Q_t(d+1) \equiv 0$, $t \ge 0$, then Proposition 2.6 yields the required result. □

We now show how (12) may be used to bound the average work done in a given time interval:

LEMMA 2.7. *If the conclusions of Lemma* 2.5 *hold, then there exists a deterministic constant* $C = C(b, T)$ *such that*

$$\mathsf{E}\left[\frac{1}{R}\int_0^R T_{s,s+t}(i)\,dt\,\middle|\,\mathscr{F}_s\right] \ge \mathsf{E}\left[\frac{1}{R}\int_0^R \Omega_{s,s+t}(i)\,dt\,\middle|\,\mathscr{F}_s\right] - C,$$
$$R, s \in \mathbb{R}_+,\ i \in I.$$

PROOF. Define $T^0_{s,s+t}(i)$ to be the busy time on $[s, s+t]$ when the queue lengths $Q_s(j)$, $1 \le j \le d$, are reduced to zero and all other properties of the network are held constant. Define $W^0_{s+t}(i)$ in a similar manner. Using Proposition 2.6 we have

$$T_{s,s+t}(i) \ge T^0_{s,s+t}(i).$$

Since we search for a lower bound on $T_{s,s+t}$, we may work with $T^0_{s,s+t}(i)$ and $W^0_{s+t}(i)$ in the proof.

In this case we have

$$W^0_{s+t}(i) \ge \Omega_{s,s+t}(i) - T^0_{s,s+t}(i)$$

and hence by (12),

$$\frac{1}{R}\int_0^R \mathsf{E}\left[\Omega_{s,s+t}(i) - T^0_{s,s+t}(i) \mid \mathscr{F}_s\right] dt \le b + \sum_{i\in I}\frac{1}{R}\int_0^T \mathsf{E}\left[W^0_{s+t}(i) \mid \mathscr{F}_s\right]^2 dt.$$

The quantity $\mathsf{E}[(W^0_{s+t}(i))|\mathscr{F}_s]^2$ is bounded by $\mathsf{E}[(\Omega_{s,s+t}(i) + Y_s(i) + 1)|\mathscr{F}_s]^2$, $t \le T$, which is uniformly bounded in $s \in \mathbb{R}_+$ by Assumptions B1, B2 and B4, so the bound above implies the result. □

PROOF OF LEMMA 2.5. We proceed with this proof by induction on d, the number of nodes in the network. For notational simplicity, we assume that

$s = 0$, as we have assumed that the upper bound on the local arrival stream rate is not dependent on the initial time s.

From the remarks following Lemma 2.5 we see that the lemma is true for $d = 1$. We will assume the result for a network with $(d - 1)$ nodes and show that the result is true for a network with d nodes.

Set

$$A_k = \{x \in \mathsf{X}: x_i \leq k \text{ for } i \leq d\}, \qquad k \in \mathbb{Z}_+.$$

Hence when $\Phi_0 \in A_k$, each buffer contains no more than k customers. In our proof we will frequently consider the process when $\Phi_0 \in A_k^c$. Under this condition we will assume without loss of generality that $Q_0(1) = \max_i Q_0(i)$. We may relabel the nodes in the system if this is not the case. Hence throughout the proof we assume that $Q_0(1)$, the buffer length at node 1, is greater than or equal to k whenever $\Phi_0 \in A_k^c$.

We see from B2, B3 and B4 that there are constants c_1 and c_2 such that

$$\begin{aligned} \mathsf{E}[W_0(1)|\mathscr{F}_0] &\geq c_1 Q_0(1), \\ Q_0(1) &\geq c_2(V_1(\Phi_0) - dS^*), \end{aligned} \tag{13}$$

where S^* is equal to $\overline{S}$ times an upper bound for the expected number of nodes a customer visits.

We will let Ξ denote the class of continuous positive functions on $\mathbb{R}_+$. Functions in Ξ are assumed deterministic, but may depend on deterministic properties of the network such as the service time distributions, or γ_i and L_i. We let ξ denote a generic element of Ξ, whose precise definition may differ at each appearance.

Under these conventions, and the assumptions of the lemma, we may obtain a downward drift for the work at the first node as follows: In a manner similar to the single queue case we may write

$$\begin{aligned} \mathsf{E}[W_t(1)|\mathscr{F}_0]^2 \leq{}& \mathsf{E}[W_0(1)|\mathscr{F}_0]^2 + 2\mathsf{E}[W_0(1)|\mathscr{F}_0](\rho_1 t + \overline{L}_1) \\ &- 2\mathsf{E}[W_0(1)|\mathscr{F}_0]\mathsf{E}[T_{0,t}(1)|\mathscr{F}_0] + \xi(t). \end{aligned} \tag{14}$$

By ignoring the time that queue 1 is busy servicing customers that arrive to node 1 in $[0, t]$, we may make the estimate

$$\mathsf{E}[T_{0,t}(1)|\mathscr{F}_0] \geq \mathsf{E}\big[\big(\tilde{W}_0(1) \wedge t\big)\big|\mathscr{F}_0\big],$$

where $\tilde{W}_0(1)$ is the work to be done by node 1 on customers that are waiting at queue 1 at time 0, not including any work done on a subsequent visit. More specifically, $\tilde{W}_0(1) = Y_0(1) + \sum_{i=1}^{Q_0(1)} S_i$, where S_i are generic i.i.d. service times at node 1.

Thus we may find k_0 sufficiently large so that $\mathsf{E}[\tilde{W}_0(1) \wedge t|\mathscr{F}_0] \geq (1 - (1 - \rho_1)/2)t$ for $\Phi_0 \in A_{k_0}^c$. The constant k_0 depends only on t and the service time distribution.

This and (14) give the bound

$$\text{(15)}\qquad \begin{aligned} \mathsf{E}[W_t(1)|\mathscr{F}_0]^2 \le \mathsf{E}[W_0(1)|\mathscr{F}_0]^2 &+ 2\mathsf{E}[W_0(1)|\mathscr{F}_0](\rho_1 t + \overline{L}_1) \\ &-(1+\rho_1)t\mathsf{E}[W_0(1)|\mathscr{F}_0] + \xi(t) \end{aligned}$$

for all $\Phi_0 \in A^c_{k_0}$.

Examining any $i \in \{2,\dots,d\}$ we find that

$$\text{(16)}\qquad \begin{aligned} \mathsf{E}[W_t(i)|\mathscr{F}_0]^2 \le \mathsf{E}[W_0(i)|\mathscr{F}_0]^2 &+ 2\mathsf{E}[W_0(i)|\mathscr{F}_0]\mathsf{E}[\Omega_{0,t}(i)|\mathscr{F}_0] \\ &- 2\mathsf{E}[T_{0,t}(i)|\mathscr{F}_0]\mathsf{E}[W_0(i)|\mathscr{F}_0] + \xi(t). \end{aligned}$$

Now, if we average both sides over t, we see that

$$\text{(17)}\qquad \begin{aligned} &\frac{1}{T}\int_0^T \mathsf{E}[W_t(i)|\mathscr{F}_0]^2\,dt \\ &\quad\le \mathsf{E}[W_0(i)|\mathscr{F}_0]^2 + 2\mathsf{E}[W_0(i)|\mathscr{F}_0] \\ &\qquad\times\frac{1}{T}\int_0^T\big(\mathsf{E}[\Omega_{0,t}(i)|\mathscr{F}_0] - \mathsf{E}[T_{0,t}(i)|\mathscr{F}_0]\big)\,dt + \xi(T). \end{aligned}$$

We now use the induction hypothesis to bound

$$\frac{1}{T}\int_0^T\big(\mathsf{E}[\Omega_{0,t}(i)|\mathscr{F}_0] - \mathsf{E}[T_{0,t}(i)|\mathscr{F}_0]\big)\,dt.$$

Consider a network possessing $(d-1)$ nodes labeled $(2,3,\dots,d)$. Customers arrive at local rate γ_j from the external arrival streams and, from an independent source, customers enter at local rate λ_1 and enter the ith queue with probability r_{1i}, $i \ge 2$, and are removed from the network with probability $r_{11} + r_{10}$.

The traffic equations for this system are

$$\text{(18)}\qquad \lambda'_i = \sum_{j=2}^{d} \lambda'_j r_{ji} + \lambda_1 r_{1i} + \gamma_i, \qquad 2 \le i \le d.$$

These are in fact the traffic equations for the original system, so

$$\text{(19)}\qquad \lambda'_i = \lambda_i, \qquad 2 \le i \le d.$$

Hence, by induction, this system is stable in the sense of Theorem 2.4.

We now return to the original system, concentrating on nodes $\{2,\dots,d\}$. The main idea of the proof is that for a fixed time horizon $[0,T]$, if k_0 is large and $\Phi_0 \in A^c_{k_0}$, then from nodes $2-d$ we obtain the $(d-1)$ node network described previously, but with the local rate $\lambda_1 r_{1i}$ increased to $\mu_1 r_{1i}$. From nodes $2-d$ we construct a specific realization of the $(d-1)$ node network described previously, which we will call the *slow subsystem*. We consider the subsystem consisting of nodes $2-d$ as the corresponding *fast subsystem*, whose total local input rate at node i is $\gamma_i + \mu_1 r_{1i}$, $2 \le i \le d$.

Consider the network consisting of nodes $2-d$ with the output of node 1 considered as an input. Any customer exiting nodes $2-d$ and entering the queue at node 1 will be considered an exiting customer by this network, which will be henceforth called the fast subsystem. We will reduce the output of node 1 to obtain an analogous slow subsystem as follows: Each time a customer completes service at node 1 a weighted coin is tossed, which is independent of the network and previous coin experiments, with probability of tails equal to ρ_1. When a "head" is obtained, this customer is removed from the system. If a "tail" is obtained, then this customer is routed to another queue or exits the system as if the coin tossing experiment did not take place.

By B4 there exists a constant L_0 such that for each $T \in \mathbb{R}_+$, there exists $k_0 = k_0(T) \in \Xi$ so large that for $\Phi_0 \in A^c_{k_0}$ and $2 \le i \le d$,

$$\lambda_1 r_{1i} t - L_0 \le \mathsf{E}\begin{bmatrix} \#\text{ customers routed from node 1 to node } i \\ \text{in the time interval } (0,t] \,|\mathscr{F}_0 \end{bmatrix} \le \lambda_1 r_{1i} t + L_0, \qquad 0 \le t \le T. \tag{20}$$

If node 1 never empties, then the slow subsystem satisfies the conditions of Theorem 2.4 and the traffic equations for the subsystem are given as in (18).

Let $\Omega^{\text{slow}}_{0,t}$ denote the work entering the slow subsystem during the time interval $(0, t]$, where the limited output stream from node 1 is considered an exogenous input. The lower bound in (20) implies that for some $c_3 > 0$,

$$\mathsf{E}\big[\Omega^{\text{slow}}_{0,t}(i)|\mathscr{F}_0\big] \ge \mathsf{E}\big[\Omega_{0,t}(i)|\mathscr{F}_0\big] - c_3, \qquad 0 \le t \le T, 2 \le i \le d, \tag{21}$$

for $\Phi_0 \in A^c_{k_0}$.

We now apply the induction hypothesis: We see from (18) that the slow subsystem satisfies the conditions of Theorem 2.4. Letting $T^{\text{slow}}_{0,t}(i)$ denote the busy time at node i for the slow subsystem, we may conclude by induction and Lemma 2.7 that

$$\mathsf{E}\left[\frac{1}{T}\int_0^T T^{\text{slow}}_{0,t}(i)\,dt|\mathscr{F}_0\right] \ge \mathsf{E}\left[\frac{1}{T}\int_0^T \Omega^{\text{slow}}_{0,t}(i)\,dt|\mathscr{F}_0\right] - c_4, \qquad 2 \le i \le d,$$

where $c_4 < \infty$ and is independent of T.

The corollary to Proposition 2.6 allows us to remove the limits on the output from node 1: From this corollary, the foregoing bound and (21) we have

$$\mathsf{E}\left[\frac{1}{T}\int_0^T T_{0,t}(i)\,dt|\mathscr{F}_0\right] \ge \mathsf{E}\left[\frac{1}{T}\int_0^T \Omega_{0,t}(i)\,dt|\mathscr{F}_0\right] - c_3 - c_4, \qquad 2 \le i \le d, \tag{22}$$

for $\Phi_0 \in A^c_{k_0}$.

Applying (22) to (17) we see that for some constant c_5,

$$\sum_{i=2}^d \frac{1}{T}\int_0^T \mathsf{E}\big[W_t(i)|\mathscr{F}_0\big]^2\,dt \le \sum_{i=2}^d \mathsf{E}\big[W_0(i)|\mathscr{F}_0\big]^2 + c_5 \sum_{i=2}^d \mathsf{E}\big[W_0(i)|\mathscr{F}_0\big] + \xi(T) \tag{23}$$

for $\Phi_0 \in A^c_{k_0}$.

We are finally ready to examine the network as a whole. From (11) we have

$$\mathsf{E}\big[V_2(\Phi_t)\,|\mathscr{F}_0\big] = \sum_{i=1}^{d} \mathsf{E}\big[W_t(i)\,|\mathscr{F}_0\big]^2 + \xi(t).$$

Combining (15), which exhibits the drift property for the work at node 1, and the bound (23) obtained for the remaining $(d-1)$ nodes gives

$$\begin{aligned}\mathsf{E}\left[\frac{1}{T}\int_0^T V_2(\Phi_t)\,dt\,|\mathscr{F}_0\right] &= \sum_{i=1}^{d} \mathsf{E}\left[\frac{1}{T}\int_0^T W_t(i)\,dt\,|\mathscr{F}_0\right]^2 dt + \xi(T)\\ &\le V_2(\Phi_0) - \frac{1}{2}(1-\rho_1)T\mathsf{E}\big[W_0(1)\,|\mathscr{F}_0\big]\\ &\quad + \bar{L}_1\mathsf{E}\big[W_0(1)\,|\mathscr{F}_0\big]\\ &\quad + c_5\sum_{i=2}^{d}\mathsf{E}\big[W_0(i)\,|\mathscr{F}_0\big] + \xi(T).\end{aligned}$$

Applying (13), we see that

$$\begin{aligned}\mathsf{E}\left[\frac{1}{T}\int_0^T V_2(\Phi_t)\,dt\,|\mathscr{F}_0\right] &\le V_2(\Phi_0) - \frac{1}{2}c_1c_2(1-\rho_1)TV_1(\Phi_0)\\ &\quad + \big(2\bar{L}_1 + c_5 + c_2\,dS^*\big)V_1(\Phi_0) + \xi(T).\end{aligned}$$

This result only holds for $\Phi_0 \in A_{k_0}^c$, where $k_0 = k_0(T) < \infty$. However, since $\mathsf{E}[(1/T)\int_0^T V_2(\Phi_t)\,dt|\mathscr{F}_0] = O(T^2)$ for $\Phi_0 \in A_{k_0}$, we may generalize this bound to all initial conditions, and we thereby obtain the result for sufficiently large T. □

3. Markovian networks. In this section we construct a Markovian state process for the network and apply recent results from the theory of continuous time Markov processes to obtain ergodicity for the state process, and hence also convergence of the mean value of most of the variables of interest for the network.

We begin with a brief review of the general results that will be needed to prove our main results.

3.1. *Continuous components and Harris ergodicity.* We present here some restricted versions of previous results [25, 26; 28]. These results will be applied in the following text to refine our stability theorem for Jackson networks.

Suppose that Ψ is a strong Markov process whose sample paths are right continuous. We assume that the process evolves on a locally compact and separable state space Y, with Borel σ-field $\mathscr{B}$(Y), and that the process Ψ is temporally homogeneous (see [11]).

The Markov process Ψ is called *Harris recurrent* if there exists a probability φ such that the following implication holds:

$$\varphi\{A\} > 0 \Rightarrow \mathsf{P}_y\{\Psi \text{ enters } A\} = 1, \qquad y \in \mathsf{Y}. \tag{24}$$

That is, the process visits any set A of positive φ-measure with probability 1.

If the process satisfies the criterion (24), then it is also Harris recurrent in the sense of [3], so that the following formally stronger condition holds: For a probability ν on Y,

$$\nu\{A\} > 0 \Rightarrow \mathsf{P}_y\left\{\int_0^\infty \mathbb{I}\{\Psi_s \in A\}\, ds = \infty\right\} = 1, \qquad y \in \mathsf{Y}.$$

Hence by a theorem in [3] (see also [11]), a Harris recurrent Markov process possesses a unique, up to scalar multiples, σ-finite invariant measure π. If the invariant measure is in fact finite, then we normalize it to be an invariant probability. In this case, Ψ is called *positive Harris recurrent*.

Let θ^s denote the shift operator on sample space, defined so that $\theta^s f(\Psi_t) = f(\Psi_{t+s})$, $t \in \mathbb{R}_+$. For a random variable Z on the sample space, we define $Z_s = \theta^s Z$. If Ψ is positive Harris recurrent, then the process $\{Z_s\}$ is strictly stationary when $\Psi_0 \sim \pi$.

Harris recurrent Markov processes enjoy a number of important ergodic properties, the most general of which is illustrated in the following theorem.

THEOREM 3.1. *If Ψ is positive Harris recurrent with invariant probability π, then for each initial condition $y \in \mathsf{Y}$ and any positive random variable Z on the sample space,*

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T Z_s\, ds = \mathsf{E}_\pi[Z] \quad a.s.\ [\mathsf{P}_y].$$

A kernel T is called a *continuous component* of a function $K\colon (\mathsf{Y}, \mathscr{B}(\mathsf{Y})) \to \mathbb{R}_+$ if:

1. For $A \in \mathscr{B}(\mathsf{Y})$, the function $T(\cdot, A)$ is lower semicontinuous.
2. For all $y \in \mathsf{Y}$ and $A \in \mathscr{B}(\mathsf{Y})$, the measure $T(y, \cdot\,)$ satisfies

$$K(y, A) \geq T(y, A).$$

The continuous component T is called *nontrivial* at y if $T(y, \mathsf{Y}) > 0$.

Suppose that a is a distribution function on $\mathbb{R}_+$ and define the Markov transition function K_a as

$$K_a := \int P^t\, da(t), \tag{25}$$

where $(P^t\colon t \in \mathbb{R}_+)$ denotes the processes' transition semigroup. If a is the increment distribution of the undelayed renewal process $\{t_k\}$, then K_a is the transition function for the Markov chain $\{\Psi_{t_k}\colon k \in \mathbb{Z}_+\}$.

We will be concerned here with continuous components of the Markov transition function K_a, as defined in (25). A process will be called a *T-process* if for some distribution a, the K_a-chain admits a continuous component T that is nontrivial for all $y \in \mathsf{Y}$. It is shown in [28] that a diffusion process is a T-process if its generator is hypoelliptic. We show in the succeeding text that the network under study in this paper is a T-process if the arrival stream satisfies certain regularity conditions.

A crucial property of T-processes is that they allow a close connection between Harris ergodicity and tightness of the distributions of a Markov process. We call the process Ψ *bounded in probability on average* (cf. [23]) if for all $y \in \mathsf{Y}$ and $\varepsilon > 0$, there exists a compact set $C \subset \mathsf{Y}$ such that

$$\limsup_{T\to\infty} \frac{1}{T}\int_0^T \mathsf{P}_y\{\Psi_s \in C^c\}\,ds \le \varepsilon.$$

This condition is implied by tightness for the family of probabilities

$$\left\{\frac{1}{T}\int_0^T P^t(y,\cdot)\,dt \colon T > 0\right\}$$

for each fixed $y \in \mathsf{Y}$.

It is shown [26] that boundedness in probability is equivalent to a generalization of positive Harris recurrence for a T-process. To obtain Harris recurrence, the following solidarity property for the process is necessary. A state y^* is called *reachable* if $\int_0^\infty P^t(y,O)\,dt > 0$ for every open set O containing y^* and every $y \in \mathsf{Y}$. If Φ is Harris recurrent, then the set of all reachable states is equal to the support of the invariant measure for the process.

The process Ψ is called *ergodic* if there exists an invariant probability π such that for every $y \in \mathsf{Y}$,

$$\lim_{t\to\infty} \|P^t(y,\cdot) - \pi\| = 0.$$

THEOREM 3.2. *Suppose that Ψ is a T-process that possesses a reachable state. Then*:

(i) *Ψ is bounded in probability on average if and only if Ψ is positive Harris recurrent.*

(ii) *Suppose that Ψ is bounded in probability on average and that for a lattice distribution a, the kernel K_a possesses an everywhere nontrivial continuous component. Then Ψ is ergodic.*

PROOF. Result (i) follows from the Doeblin decomposition [34, 26] and the observation that the closure of distinct Harris sets is disjoint for a T-process. Theorem 7.1 of [26] establishes the limit theorem for all initial conditions. □

Under the conditions of Theorem 3.2 the expectation of bounded functions of the process converge to a steady state value for all initial conditions. To generalize this result to unbounded functions we need the concept of the

f-norm $\|\mu\|_f$. For any positive measurable function $f \geq 1$ and any signed measure μ on $\mathscr{B}(\mathsf{X})$ we write

$$\|\mu\|_f = \sup_{|g| \leq f} |\mu(g)|.$$

Note that the total variation norm $\|\mu\|$ is $\|\mu\|_f$ in the special case where $f \equiv 1$. We call Ψ f-ergodic if:

1. Ψ is ergodic with invariant probability π.
2. $\int f \, d\pi < \infty$.
3. For each initial condition $y \in \mathsf{Y}$,

$$\lim_{t \to \infty} \|P^t(y, \cdot) - \pi\|_f = 0.$$

The following result is taken from [26].

THEOREM 3.3. *Suppose that for a lattice distribution a, the kernel K_a possesses an everywhere nontrivial continuous component, and suppose that a reachable state exists. Let $f \colon \mathsf{Y} \to [1, \infty)$ satisfy for constants $\delta > 0$, $c_\delta < \infty$,*

$$P^s f \leq c_\delta f, \qquad 0 \leq s \leq \delta.$$

Suppose that A is compact, $t > 0$, and that the expectation

$$\mathsf{E}_y\left[\int_0^{t + \theta^t \tau_A} f(\Psi_s)\, ds\right] \tag{26}$$

is everywhere finite and uniformly bounded for $y \in A$. Then Ψ is f-ergodic.

The quantity $t + \theta^t \tau_A$ is equal to the first time $s \geq t$ that the process enters the set A. The event may not occur, in which case we set $t + \theta^t \tau_A = \infty$. A simple method for estimating the expectation (26), based upon the drift property for the network obtained in Lemma 2.5, is illustrated in the next section. An analogous technique based upon the infinitesimal generator for the process is developed in [28].

We now show how all of these results may be applied to the process under study in this paper.

3.2. *Markovian Jackson networks.* Here we prove what is perhaps the most important result of the paper. First we construct a Markovian state process Ψ for the network, and then we proceed to establish the conditions of Theorem 3.3.

Suppose that $N(s)$ is a counting process for a delayed renewal process, with deterministic delay $F_0 \in \mathbb{R}_+$. We let t_k denote the kth jump of $N(s)$, $k \geq 1$, which are also the times at which an arrival to the network occurs. Suppose that there exists a separate, Bernoulli routing policy that is independent of $N(s)$ and independent of the routing in the network and the service times, such that at time t_k, a single customer is routed to node i with probability p_i. We let Γ_k denote the node that is chosen at time t_k and let $\{N_i(s): i \in I\}$ denote the resulting arrival streams.

We define $\Psi_t = \binom{\Phi_t}{F_t}$, $t \in \mathbb{R}_+$, where $\{F_t\}$ denotes the forward process for the renewal process. Under the conditions of this section, the process Ψ serves as a Markovian state process for the network:

LEMMA 3.4. *The stochastic process Ψ is a temporally homogeneous, strong Markov process with right continuous sample paths, whose state space* Y *is equal to* $\mathsf{X} \times \mathbb{R}_+$.

The proof essentially follows the proof of Proposition 1.5, page 108 of [2], or see [7, 8].

For $k \geq 2$ we define $v_k = t_k - t_{k-1}$. The random variables $\{v_k: k \geq 2\}$ are i.i.d., with common distribution denoted ν. We let γ^{-1} denote the common mean of v_k, which we assume is finite. Under this condition the forward process itself is bounded in probability on average and possesses a unique invariant probability (see page 142 of [2]).

For $i \in I$, let $\gamma_i := p_i \gamma$. We let

$$\mathscr{F}_s := \sigma\{\Phi_r, F_r, \Gamma_k : r \leq s, t_k \leq s\}.$$

Applying Theorem 2.4(iii) on page 113 of [2] we see that these arrival streams satisfy the assumptions introduced in Section 2:

LEMMA 3.5. *The arrival streams constructed in this section satisfy for all s, t and i,*

$$\mathsf{E}[N_i(s+t) - N_i(s) | \mathscr{F}_s] \leq \gamma_i t + C_i,$$

where C_i is a deterministic constant.

This combined with Theorem 2.4 and the stability properties of the forward process immediately imply the following lemma.

LEMMA 3.6. *If the load condition* (8) *holds for the Markovian network, then the process Ψ is bounded in probability on average.*

By strengthening slightly the assumptions on the arrival stream we can construct a continuous component, which will allow us to say far more. Suppose that the interarrival distribution ν is not supported on a bounded subset of $\mathbb{R}_+$, and that ν is spread out. These conditions are listed here:

(27) The distribution ν is unbounded, so that $\nu\{[L, \infty)\} > 0$ for all $L > 0$, and for some $k_0 \geq 2$, the $(k_0 - 1)$-fold convolution $\nu^{(k_0-1)*}$ is nonsingular with respect to Lebesgue measure.

It is well known that the spread-out condition is equivalent to Harris ergodicity of the forward process (cf. Corollary 1.5, page 142 of [2], or [1]). Under this condition we may construct a continuous component for the process:

LEMMA 3.7. *If the interarrival distribution satisfies* (27), *then for a lattice distribution* a, *the kernel* K_a *possesses an everywhere nontrivial continuous component, and the state* $0 \in \mathsf{Y}$ *is reachable for the Markov process* Ψ.

The proof of Lemma 3.7 is included in the following text.

Using Lemma 3.7 and Lemma 2.5 we may prove the following ergodic theorem for Jackson networks:

THEOREM 3.8. *If the load condition* (8) *holds for the network, and if the interarrival distribution satisfies* (27), *then* Ψ *is f-ergodic, with* $f(\Psi) = f\binom{\Phi}{F} = 1 + V_1(\Phi)$. *In particular, we have*

$$\lim_{t\to\infty} \mathsf{E}_y[Q_t(i)] = \mathsf{E}_\pi[Q_0(i)],$$

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathsf{E}_y[Q_t(i)]\,dt = \mathsf{E}_\pi[Q_0(i)] \quad a.s.$$

for every initial condition $y \in \mathsf{Y}$.

Sigman [32] also obtains Harris ergodicity for Markovian networks where the analysis is largely restricted to the two node case. The generality of the assumptions and proof technique of [32] rule out the possibility of obtaining bounds on the expected work in the system.

Borovkov [4] also obtains Harris ergodicity for generalized Jackson networks. However, there is a difficulty with this work in that hypothesized arrival rates of customers to each node in the network are employed before any stability result has been established.

The proof of Theorem 3.8, given the foregoing results, is relatively straightforward.

PROOF OF THEOREM 3.8. Under the conditions of the theorem we have, from Lemma 2.5 and the Markovian hypotheses, that for all large $T > 0$ there exists $b < \infty$ such that

$$\frac{1}{T}\int_0^T P^s V_2(y)\,ds \le V_2(y) - V_1(y) + b, \qquad y \in \mathsf{Y}. \tag{28}$$

Define $B_i(\Psi_t) = B_t(i)$. From B4, we find that for suitably large T, there exists $c < \infty$ such that

$$\frac{1}{T}\int_0^T \mathsf{E}_y[B_i(\Psi_s)]\,ds \le \frac{1}{2}B_i(\Psi_0) + c, \qquad y \in \mathsf{Y}. \tag{29}$$

Define $V_3(\Phi_t) = V_2(\Phi_t) + \sum_{i=1}^d B_i(\Phi_t)$. From the preceding bounds we see that we can find a compact set $A \subset \mathsf{Y}$ suitably large so that

$$\frac{1}{T}\int_0^T P^s V_3(y)\,ds \le V_3(y) - \frac{1}{2}f(y), \qquad y \in A^c, \tag{30}$$

where $f(y) = V_1(y) + 1$, $y \in \mathsf{Y}$. This is as a result of considering three cases. If $y \in A^c$ due to a large initial queue size, then (30) follows from (28). Similarly, if $y \in A^c$ due to a large initial backward recurrence time, then (30) follows from (29). Otherwise, (30) follows from the fact that there are no new arrivals to the system.

Let τ_A denote the first entrance time to A. On substituting Ψ_t for y in (30), integrating from $t = 0$ to $\tau_A \wedge n$ and taking expectations, we have for any $n \in \mathbb{Z}_+$,

$$\begin{aligned}\frac{1}{T}\mathsf{E}_y\biggl[\int_0^{\tau_A\wedge n}\int_0^T P^s V_3(\Psi_t)\,ds\,dt\biggr]\\ \le \mathsf{E}_y\biggl[\int_0^{\tau_A\wedge n} V_3(\Psi_t)\,dt\biggr] - \frac{1}{2}\mathsf{E}_y\biggl[\int_0^{\tau_A\wedge n} f(\Psi_t)\,dt\biggr].\end{aligned}$$

By the Markov property, the left-hand side of this inequality is greater than

$$\mathsf{E}_y\biggl[\int_0^{\tau_A\wedge n} V_3(\Psi_t)\,dt\biggr] - \mathsf{E}_y\biggl[\int_0^T V_3(\Psi_t)\,dt\biggr],$$

which immediately gives the bound

$$\tfrac{1}{2}\mathsf{E}_y\biggl[\int_0^{\tau_A\wedge n} f(\Psi_t)\,dt\biggr] \le \mathsf{E}_y\biggl[\int_0^T V_3(\Psi_t)\,dt\biggr].$$

Letting $n \to \infty$ we have by the monotone convergence theorem,

$$\mathsf{E}_y\biggl[\int_0^{\tau_A} f(\Psi_t)\,dt\biggr] \le U(y), \qquad y \in \mathsf{X}, \tag{31}$$

where $U(y) := 2\mathsf{E}_y[\int_0^T V_3(\Psi_t)\,dt]$.

From (31) we also have for any $s \in \mathbb{R}_+$,

$$\int P^s(y,dz)\mathsf{E}_z\biggl[\int_0^{\tau_A} f(\Psi_t)\,dt\biggr] \le P^s U(y) = 2\mathsf{E}_y\biggl[\int_s^{s+T} V_3(\Psi_t)\,dt\biggr], \tag{32}$$

which shows that

$$\begin{aligned}\mathsf{E}_y\biggl[\int_0^{s+\theta^s\tau_A} f(\Psi_t)\,dt\biggr] &= \mathsf{E}_y\biggl[\int_0^s f(\Psi_t)\,dt\biggr]\\ &\quad + \int P^s(y,dz)\mathsf{E}_z\biggl[\int_0^{\tau_A} f(\Psi_t)\,dt\biggr]\end{aligned} \tag{33}$$

is everywhere finite and uniformly bounded for y in compact subsets of Y. In particular, (33) is uniformly bounded on A.

Finally, from (10) we may find $K < \infty$ such that

$$P^s f(y) \le f(y) + Ks, \qquad y \in \mathsf{Y},\ s \in \mathbb{R}_+ \tag{34}$$

Hence from Lemma 3.7 and (32), (33) and (34) we see that the conditions of Theorem 3.3 are satisfied, which completes the proof of the theorem. □

PROOF OF LEMMA 3.7. The random variables $\{v_k\}$ are i.i.d. with common distribution ν, and by assumption there exists $j_0 = k_0 - 1 \geq 1$ such that ν^{*j_0} is nonsingular with respect to Lebesgue measure. That is, there exists a positive function p on $\mathbb{R}_+$ with

$$\nu^{*j_0}(dx) \geq p(x)\,dx \quad \text{and} \quad \int_0^\infty p(x)\,dx > 0. \tag{35}$$

For arbitrary $l \in \mathbb{Z}_+$, let R_l denote the bounded open rectangle in Y defined as

$$R_l = \{0,\dots,l\}^d \times [0,l)^d \times [0,l).$$

We first construct a continuous component that is nontrivial on R_l.

We estimate, for large n, the probability $P^n(\Psi_0,\{0\}\times A)$ for $\Psi_0 \in R_l$, $A \in \mathscr{B}(\mathbb{R}_+)$ as follows: For large n we have,

$$\begin{aligned} P^n(\Psi_0,\{0\}\times A) &\geq P_{\Psi_0}\left\{\Phi_n = 0,\, F_n \in A,\, \sum_{i=2}^{k_0} \nu_i \leq \frac{n}{2},\, v_{k_0+1} \geq 2n\right\} \\ &\geq \mathsf{P}\left\{E_{L,l}, F_n \in A,\, \sum_{i=2}^{k_0} \nu_i \leq \frac{n}{2},\, v_{k_0+1} \geq 2n\right\}, \end{aligned} \tag{36}$$

where $E_{L,l}$ denotes the event that no customer is routed to the same queue twice and that each of the first $d(dl + k_0)$ services take no more than L units of time. The quantity $d(dl + k_0)$ is an upper bound on the number of services required for the customers initially waiting in the queues, and for the k_0 customers that arrive before time n. L is chosen large enough so that $\varepsilon_0 := \mathsf{P}(E_{L,l}) > 0$.

Since $E_{L,l}$ is independent of $\{F_t\colon 0 \leq t < \infty\}$ we have from (36),

$$\begin{aligned} &P^n(\Psi_0,\{0\}\times A) \\ &\quad\geq \varepsilon_0 \mathsf{P}\left\{F_n \in A,\, \sum_{i=2}^{k_0} \nu_i \leq \frac{n}{2},\, v_{k_0+1} \geq 2n\right\} \\ &\quad= \varepsilon_0 \mathsf{P}\left\{\left(v_{k_0+1} + \sum_{i=2}^{k_0} \nu_i - n + F_0\right) \in A,\, \sum_{i=2}^{k_0} \nu_i \leq \frac{n}{2},\, v_{k_0+1} \geq 2n\right\} \\ &\quad= \varepsilon_0 \int_0^\infty \int_0^\infty \mathbb{I}\{(r+s-n+F_0) \in A\}\mathbb{I}\left\{s \leq \frac{n}{2}\right\}\mathbb{I}\{r \geq 2n\}\,\nu^{j_0*}(ds)\,\nu(dr) \\ &\quad\geq \varepsilon_0 \int_0^\infty \int_0^\infty \mathbb{I}\{(r+s-n+F_0) \in A\}\mathbb{I}\left\{s \leq \frac{n}{2}\right\}\mathbb{I}\{r \geq 2n\}\,p(s)\,ds\,\nu(dr). \end{aligned}$$

Define

$$\begin{aligned} T_l'(F_0, A) := \varepsilon_0 &\int_0^\infty \int_0^\infty \mathbb{I}\{(r+s-n+F_0) \in A\} \\ &\times \mathbb{I}\left\{s \leq \frac{n}{2}\right\}\mathbb{I}\{r \geq 2n\}\,p(s)\,ds\,\nu(dr). \end{aligned}$$

By construction we have for any set $R \times A \in \mathscr{B}(\mathsf{X} \times \mathbb{R}_+)$ and any $\Psi_0 \in R_l$,

$$P^n(\Psi_0, R \times A) \geq \delta_0\{R\} T_l'(F_0, A),$$

where δ_0 is the unit mass concentrated on $0 \in \mathsf{X}$. It is easy to see that T_l' is a continuous function of F_0 and by construction, $T_l'(F_0, \mathsf{X}) > 0$ for all $F_0 < l$. Hence the kernel

$$T_l(\Psi_0, R \times A) := \mathbb{I}\{\Psi_0 \in R_l\} \delta_0\{R\} T_l'(F_0, A)$$

is a continuous component of P^n, which is nontrivial for $\Psi_0 \in R_l$.

We now construct a continuous component that is everywhere nontrivial. Letting n_l denote an integer time at which P^{n_l} admits a continuous component T_l that is nontrivial on R_l, we define the distribution a on $\mathbb{R}_+$ as

$$a = \sum_{l=1}^{\infty} 2^{-l} \delta_{n_l}$$

and the kernel T as

$$T = \sum_{l=1}^{\infty} 2^{-l} T_l.$$

Then T is a continuous component of K_a that is everywhere nontrivial.

It is also easy to see that $y^* = 0 \in \mathsf{Y}$ is reachable, and this concludes the proof. □

4. Conclusion. We have presented a proof of stability for generalized Jackson networks in a non-Markovian seting. In the special case of Markovian networks we have made use of previous results [26, 28] to show that the expected value of the queue lengths, work in the system and customer delay all converge to their steady state values.

The appeal of our methodology is that it is conceptually simple, and the burden of constructing a Markovian model may be lifted. We believe that these methods will be applicable to more complex models that arise in areas such as manufacturing and communications networks.

Work has been done on strengthening these results to exponential ergodicity. Spieksma and Tweedie [33] (and generalized in [27]) show that if for some function V, a compact set C and constants T and b the following drift property is satisfied,

$$\mathsf{E}[V(\Phi_{s+T}) | \mathscr{F}_s] \leq V(\Phi_s) - 1 + b\mathbb{I}\{\Phi_s \in C\}$$

and if the increments $V(\Phi_{s+T}) - V(\Phi_s)$ are suitably bounded, then the test function $V^*(x) = \exp(\delta V(x))$ satisfies

$$\mathsf{E}[V_{t+s}^* | \mathscr{F}_s] \leq \lambda V_s^* + b,$$

where $\lambda < 1$. These conditions are satisfied for this model with

$$V(x) = V_1(x) + F(x) + 1 = \mathsf{E}_x[W_0] + F(x) + 1 \tag{37}$$

whenever the distribution of the interarrival times has exponentially decaying tails, and from this we get the following theorem.

THEOREM 4.1. *If the load condition* (8) *holds for the network, and if the interarrival distribution possesses a moment generating function defined in a neighborhood of the origin, then* Ψ *is geometrically ergodic. Furthermore, letting* π *denote the invariant probability for the process, we have*

$$\|P^t(y,\cdot) - \pi\{\cdot\}\|_{V^*} \le RV^*(y)\rho^t, \qquad y \in \mathsf{Y},$$

with $R < \infty$, $\rho < 1$ *and* V *given by* (37).

The proof follows from Theorem 16.0.1 of [27] as generalized to continuous time processes in [9].

Fayolle, Malyshev, Mensikov and Sidorenko [10] have previously established geometric ergodicity in the special case of Poisson arrivals and exponential service times.

It may be shown that a functional central limit theorem holds for the Markov process, allowing a suitably normalized version of the queue lengths to be approximated by a Brownian motion. The reader is referred to [12] for a development of this result.

Of course, an important aspect of network analysis is performance. A topic of future interest would be deriving tight bounds on the customer delay and the queue lengths for the network. It is our hope that these results will lend themselves to such calculations.

Acknowledgments. The authors are grateful to Ayalvadi Ganesh, who pointed out an error in an earlier version of this paper, leading us to a simpler proof of Lemma 2.5. We would also like to thank Karl Sigman for the example indicating the difference between local and long-term arrival rates contained in Section 2.

REFERENCES

[1] ARJAS, E., NUMMELIN, E. and TWEEDIE, R. L. (1978). Uniform limit, theorems for non-singular renewal and Markov renewal processes. *J. Appl. Probab.* **15** 112–125.

[2] ASMUSSEN, S. (1987). *Applied Probability and Queues*. Wiley, New York.

[3] AZÉMA, J., KAPLAN-DUFLO, M. and REVUZ, D. (1967). Measure invariante sur les classes récurrentes des processus de Markov. *Z. Wahrsch. Verw. Gebiete* **8** 157–181.

[4] BOROVKOV, A. A. (1986). Limit theorems for queueing networks. *Theory Probab. Appl.* **31** 413–427.

[5] CRUZ, R. L. (1991). A calculus for network delay, part I: Network elements in isolation. *IEEE Trans. Inform. Theory* **37** 114–131.

[6] CRUZ, R. L. (1991). A calculus for network delay, part II: Network analysis. *IEEE Trans. Inform. Theory* **37** 132–141.

[7] DAVIS, M. H. A. (1984). Piecewise-deterministic Markov processes: A general class of nondiffusion stochastic models. *J. Roy. Statist. Soc. Ser. B* **46** 353–388.

[8] DAVIS, M. H. A. (1992). *Markov Models and Optimization*. Chapman and Hall, London.

[9] DOWN, D., MEYN, S. P. and TWEEDIE, R. L. (1993). Exponential and uniform ergodicity of Markov processes. Unpublished manuscript.

[10] FAYOLLE, G., MALYSHEV, V. A., MENSIKOV, M. V. and SIDORENKO, A. F. (1991). Lyapunov functions for Jackson networks. Technical Report 1380, INRIA Rocquencourt.

[11] GETOOR, R. K. (1979). Transience and recurrence of Markov processes. In *Séminaire de Probabilités XIV* (J. Azéma and M. Yor, eds.) 397–409. Springer, Berlin.
[12] GLYNN, P. W. and MEYN, S. P. (1994). A Lyapunov bound for solutions of Poisson's equation. *Ann. Probab.* To appear.
[13] JACKSON, J. R. (1963). Jobshop-like queueing systems. *Management Sci.* **10** 131–142.
[14] KASPI, H. and MANDELBAUM, A. (1992). Regenerative closed queueing networks. *Stochastics.* To appear.
[15] KELLY, F. P. (1987). *Reversibility and Stochastic Networks.* Wiley, New York.
[16] KINGMAN, J. F. C. (1961). The ergodic behaviour of random walks. *Biometrika* **48** 391–396.
[17] KONSTANTOPOULOS, P. and WALRAND, J. (1989). Stationarity and stability of fork–join networks. *J. Appl. Probab.* **26** 604–614.
[18] KONSTANTOPOULOS, P. and WALRAND, J. (1990). On the ergodicity of networks of $\cdot/GI/1/N$ queues. *Adv. in Appl. Probab.* **22** 263–267.
[19] KUMAR, P. R. and MEYN, S. P. (1993). Stability of queueing networks and scheduling policies. Presented at the 32nd IEEE Conference on Decision and Control.
[20] KUMAR, P. R. and SEIDMAN, T. I. (1990). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automat. Control* **AC-35** 289–298.
[21] KUSHNER, H. J. (1967). *Stochastic Stability and Control.* Academic, New York.
[22] LU, S. H. and KUMAR, P. R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automat. Control* **36** 1406–1416.
[23] MEYN, S. P. and CAINES, P. E. (1991). Asymptotic behavior of stochastic systems processing Markovian realizations. *SIAM J. Control Optim.* **29** 535–561.
[24] MEYN, S. P. and DOWN, D. (1991). Bounding delays in open Jackson networks. In *Proceedings of the 25th Conference on Information Sciences and Systems* 408–416. Johns Hopkins Univ. Press.
[25] MEYN, S. P. and TWEEDIE, R. L. (1992). Stability of Markovian processes I: Discrete time chains. *Adv. in Appl. Probab.* **24** 542–574.
[26] MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes II: Continuous time processes and sampled chains. *Adv. in Appl. Probab.* **25** 487–517.
[27] MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer, London.
[28] MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes III: Foster–Lyapunov criteria for continuous time processes. *Adv. Appl. Probab.* **25** 518–548.
[29] NUMMELIN, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge Univ. Press.
[30] PERKINS, J. R. and KUMAR, P. R. (1989). Stable distributed real-time scheduling of flexible manufacturing/assembly/disassembly systems. *IEEE Trans. Automat. Control* **AC-34** 139–148.
[31] KUMAR, S. and KUMAR, P. R. (1992). Performance bounds for queueing networks and scheduling policies. Technical report, Coordinated Science Laboratory, Univ. Illinois, Urbana.
[32] SIGMAN, K. (1990). The stability of open queueing networks. *Stochastic Process. Appl.* **35** 11–25.
[33] SPIEKSMA, F. M. and TWEEDIE, R. L. (1992). Strengthening ergodicity to geometric ergodicity for Markov chains. *Stochastic Models* **10**. To appear.
[34] TUOMINEN, P. and TWEEDIE, R. L. (1979). Markov chains with continuous components. *Proc. London Math. Soc.* **38** 89–114.

COORDINATED SCIENCE LABORATORY
DECISION AND CONTROL LABORATORY
UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN
1308 WEST MAIN STREET
URBANA, ILLINOIS 61801