On Exponential Ergodicity of Multiclass Queueing Networks

On Exponential Ergodicity of Multiclass Queueing Networks

David Gamarnik* MIT

Sean Meyn† UIUC

February 2, 2008

Abstract

One of the key performance measures in queueing systems is the exponential decay rate of the steady-state tail probabilities of the queue lengths. It is known that if a corresponding fluid model is stable and the stochastic primitives have finite moments, then the queue lengths also have finite moments, so that the tail probability $\mathbb{P}(\cdot > s)$ decays faster than s^{-n} for any n. It is natural to conjecture that the decay rate is in fact exponential.

In this paper an example is constructed to demonstrate that this conjecture is false. For a specific stationary policy applied to a network with exponentially distributed interarrival and service times it is shown that the corresponding fluid limit model is stable, but the tail probability for the buffer length decays slower than $s^{-\log s}$.

1 Introduction.

A key performance measures in queueing models is the decay rate of the queue length distribution in steady state [9, 8, 6, 13]. Except in trivial cases, the decay rate is at best exponential for networks with a finite number of servers. Moreover, an exponential decay rate can be verified in many queueing models either by direct probabilistic arguments such as Kingman's classical bound for the G/G/1 queueing system, using Lyapunov function type arguments [15, 11, 7] (see in particular Sections 16.3 and 16.4 of [16]), using large deviations techniques [19], and even specialized techniques based on a fluid limit model [14, Theorem 4].

*MIT Sloan School of Management, Cambridge, MA, 02139 e-mail: gamarnik@mit.edu

†Department of Electrical and Computer Engineering and the Coordinated Sciences Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, e-mail: meyn@control.csl.uiuc.edu

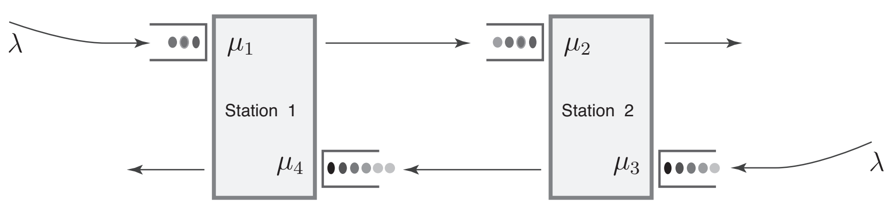

Figure 1: Rybko-Stolyar network

Except in special cases, such as single-class queueing networks [7], verifying the existence of an exponential tail is non-trivial precisely because verifying stability of a multi-class network is no longer straightforward. It is now well known that the standard load condition $\rho_\bullet < 1$ is not sufficient for stability. This was demonstrated in the two seminal papers of Kumar and Seidman [10] and Rybko and Stolyar [18], based on the network depicted in Figure 1, henceforth called the *KSRS network*.

Motivated in part by these examples, over the past ten years new methods have been developed to verify stability. The most general techniques are based on the fluid limit model (recalled below equation (5)) starting with the work of Malyšev and Men′šikov [12] and Stolyar [20]. This was extended to a broad class of multiclass networks by Dai [2].

Dai showed that stability of the fluid limit model together with some mild conditions on the network implies positive Harris recurrence for a Markovian state process, which implies in particular the existence of a unique steady state for the underlying queueing network. This result was extended in Dai and Meyn [3] where it is shown that the queue lengths have finite moments in steady state up to order p if the stochastic primitives of the network (interarrival and service time distributions) have finite moments up to $p + 1$.

As a direct implication of the main result of [3], if the stochastic primitives have an exponential moment, and if the fluid limit model is stable, then the decay rate of the queue length distribution in steady state is faster than than any polynomial, in the sense that, for any $p \geq 1$,

$$\lim_{s\to\infty} r(s)\mathbb{P}\{\|Q(0)\| \geq s\} = 0, \tag{1}$$

where $r(s) = s^p$, and the probability is with respect to a stationary version of the queue.

It is then natural to conjecture that an exponential bound holds, so that (1) holds with $r(s) = e^{\theta s}$ for some $\theta > 0$, provided the stochastic primitives possess exponentially decaying tails. The purpose of this paper is to construct a particular stationary policy and a particular network to demonstrate that this conclusion in fact does not hold. Moreover, a polynomial rate is about the best that can be attained: In the example it is found that (1) cannot hold for any sequence $\{r(s)\}$ satisfying $\liminf_{s\to\infty} r(s)/s^{\log(s)} > 0$.

The example is a particular instance of the KSRS model in which the interarrival and service times have exponential distributions. The scheduling policy is stationary, and the corresponding fluid limit model is stable, yet the queue length process has heavy tails in steady-state. The policy is based on carefully randomizing between a stable policy, and the unstable policy introduced by Rybko and Stolyar.

The remainder of the paper is organized as follows. In the following section we describe the model and the main result. In Section 3 we present a short proof of the earlier result of [3] in the special setting of this paper. The construction of the counterexample is provided in Section 4. The details of the proof of the main result are contained in Sections 5 and 6. Some technical arguments are placed in an appendix.

We close the introduction with some notational conventions: The ith unit vector in $\mathbb{R}^N$ is denoted e_i for $1 \leq i \leq N$. The L_1-norm, always denoted $\|\cdot\|$, is defined by $\|a\| = \sum_{1 \leq i \leq N} |a_i|$ for $a \in \mathbb{R}^N$. Given two positive real valued functions $f(x), g(x)$, the notation $f = O(g)$ stands for $f(x) \leq Cg(x)$ for some constant $C > 0$ and all $x \geq 0$. Similarly, $f = \Omega(g)$ means $f(x) \geq Cg(x)$ for all $x \geq 0$ and $f = \Theta(g)$ means $f = O(g)$ and $f = \Omega(g)$ at the same time. Given a Markov chain or a Markov process $\boldsymbol{Z}$ defined on a space $\mathcal{Z}$ and given a probability measure ν on $\mathcal{Z}$, we let $\mathbb{P}_\nu(Z(t))$ denote the law of $Z(t)$ initialized by have $Z(0)$ distributed according to ν. Specifically, for every $x \in \mathcal{Z}$, $\mathbb{P}_x(Z(t))$ is the law of $Z(t)$ conditioned on $Z(0) = x$. The notations $\mathbb{E}_\nu[\cdot], \mathbb{E}_x[\cdot]$ have similar meaning corresponding to the expectation operator.

2 Model description and main result

The model and the definitions of this paper follow closely those of [3]. We consider a multiclass queueing network consisting of J servers denoted simply by $1, 2, \ldots, J$. Each customer class is associated with an exogenous arrival process which is assumed to be a renewal process with rate λ_i. Here N denotes the number of classes. It is possible that $\lambda_i = 0$ for some of the classes, namely, no external arrival is associated with this class. We let $\lambda = (\lambda_i), 1 \leq i \leq N$. Each server is unit speed and can serve customers from a fixed set of customer classes $i = 1, 2, \ldots, N$. The classes are associated with servers using the constituency matrix C, where $C_{ij} = 1$ if class i is served by server j, and $C_{ij} = 0$ otherwise. It is assumed that each class is served by exactly one server, but the same server can be associated with many classes. Each class is associated with a buffer at which the jobs are queued waiting for service. The queue length corresponding to the jobs in buffer i at time t is denoted by $Q_i(t)$, and $Q(t)$ denotes the corresponding N-dimensional buffer.

It is assumed that routing is deterministic: The routing matrix R has entries equal to zero or one. Upon service completion, a job in class i proceeds to buffer i_+, where i_+ denotes the index satisfying $R_{i\,i_+} = 1$, provided such an index exist. In this case buffer i is called an *internal buffer*. If no such index exists, then this is an *exit buffer*, and the completed job leaves the network. The network is assumed to be open, so that $R^N = 0$.

For the purpose of building a counterexample we restrict to a Markovian model: The arrival processes are assumed Poisson, and the service times have exponential distributions. It is also convenient to *relax* the assumptions above and allow infinite rates for service at certain queues. That is, the corresponding service time is *zero.* For this reason it is necessary to show that the main result of [3] can be extended to this setting.

We can express the evolution of the queue length process as

$$
\begin{aligned}
Q(t) &= Q(0) + \sum_{i,j=1}^{N} [-e_i + R_{ij}] D_i(t) + A(t) \\
&= Q(0) + [-I + R^{\mathrm{T}}] D(t) + A(t),
\end{aligned}
\tag{2}
$$

where $\boldsymbol{A}_i$ is the cumulative arrival process to buffer i, and $\boldsymbol{D}_i$ is the cumulative departure process from buffer i. The second equation is in vector form with R equal to the routing matrix, and $D(t), A(t)$ the N-dimensional vectors of departures and arrivals.

If the service rate μ_i is finite, then the departure process can be expressed,

$$D_i(t) = S_i(Z_i(t)), \qquad i = 1, \dots, N, \ t \geq 0,$$

where $\boldsymbol{S}_i$ is a Poisson process with rate μ_i, and $Z_i(t)$ is the cumulative busy time at buffer i. All of these processes are assumed right continuous.

The following assumptions are imposed on the policy that determines $\boldsymbol{Z}$: It is assumed throughout the paper that the policy is *stationary.* In this setting, for buffers with finite service rate this means that $\frac{d}{dt} Z_i(t)$ is piecewise constant, and when the derivative exists it can be expressed as a fixed function of the queue-length process, $\frac{d}{dt} Z_i(t) = \phi_i(Q(t))$. Moreover, for each i satisfying $\mu_i = \infty$ there is a fixed set of values $\Xi_i \subset \mathbb{Z}_+^N$ such that the contents of buffer i are drained at the instant $Q(t) \in \Xi_i$.

We also considered randomized stationary policies. In this case ϕ_i is a randomized function of $Q(t)$, and the draining of a buffer with infinite service rate occurs with some probability depending upon the particular value of $Q(t) \in \Xi_i$ observed at time t.

It is assumed that the policy is *non-idling*: For each station we have $\sum \frac{d}{dt} Z_i(t) = 1$ when $\sum Q_i(t) > 0$, where the sum is over i at the given station.

Throughout much of the paper we restrict to the KSRS model in which the routing and constituency matrix are expressed,

$$
R^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad C = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}
$$

The network is symmetric: The two non-null arrival processes are independent Poisson processes with rates $\lambda_2 = \lambda_4 > 0$, and the service rates at buffers 2 and 4 are finite and equal, $\mu_2 = \mu_4 < \infty$. The service rates at buffers 1 and 3 are infinite. Hence, for example, if at any moment priority is given to buffer 1, then all of the contents pass instantly to buffer 2.

The transition semigroup for $\boldsymbol{Q}$ is denoted,

$$P^t(x, A) = \mathbb{P}\{Q(t) \in A \mid Q(0) = x\}, \qquad x \in \mathbb{Z}_+^n,\ t \geq 0, A \subset \mathbb{Z}_+^N,$$

and $\mathbb{P}_x(\,\cdot\,)$ the probability law corresponding to the initial state $Q(0) = x$. A probability measure π on the state space $\mathbb{Z}_+^N$ is invariant if $\boldsymbol{Q}$ is a stationary process when initialized using π. This is equivalently expressed by the invariance equations,

$$\pi(y) = \sum_x \pi(x) P^t(x, y), \qquad y \in \mathbb{Z}_+^N,\ t \geq 0.$$

We say that π has an *exponential tail* if for some $\theta > 0$,

$$\sum_x \pi(x) e^{\theta \|x\|} < \infty.$$

The Markov process is called *exponentially ergodic* if π exists, and for some $\theta > 0$ and each x, y,

$$\lim_{t \to \infty} e^{\theta t} |\mathbb{P}_x(Q(t) = y) - \pi(y)| = 0. \tag{3}$$

Exponential ergodicity implies an exponential tail. The proof of Proposition 2.1 is provided in Section 3.

Proposition 2.1. *Consider the network* (2) *in which $\mu_i < \infty$ for each exit buffer. Assume that the network is controlled using a stationary policy. If $\boldsymbol{Q}$ is exponentially ergodic, then it has an exponential tail.*

We now construct the *fluid limit model* associated with the network (2). To emphasize the dependence on the initial state $Q(0) = x$ we denote the queue and allocation trajectories by $Q(\cdot, x), Z(\cdot, x)$. The scaled initial condition is defined for $\kappa > 0$ by,

$$x^\kappa := \frac{1}{\kappa} \lfloor \kappa x \rfloor, \qquad x \in \mathbb{R}_+^N, \tag{4}$$

and the scaled processes are defined for $t \geq 0$ via,

$$q^\kappa(t; x^\kappa) := \frac{1}{\kappa} Q(\kappa t; \kappa x^\kappa), \quad z^\kappa(t; x^\kappa) := \frac{1}{\kappa} Z(\kappa t; \kappa x^\kappa). \tag{5}$$

Observe that $x^\kappa \in \mathbb{R}_+^N$ satisfies $\kappa x^\kappa \in \mathbb{Z}_+^N$ for each κ. For each $x \in \mathbb{Z}_+^N$ and $\omega \in \Omega$ we let $\mathcal{L}_x(\omega)$ denote the set of all possible fluid limits,

$$\mathcal{L}_x(\omega) = \Big\{ \text{ u.o.c. subsequential limits of } \{q^\kappa(t;x^\kappa,\omega),\ z^\kappa(t;x^\kappa,\omega)\} \Big\},$$

where 'u.o.c.' means that convergence is uniform on compact time intervals as $\kappa_i \to \infty$ for some subsequence $\{\kappa_i\}$. The *fluid limit model* is the union $\mathcal{L} := \cup_{x\in\mathbb{Z}^n_+} \mathcal{L}_x$. It is clear that any fluid limit must satisfy the fluid model equation,

$$q(t) = q(0) + [-I + R^{\mathrm{T}}]z(t) + \lambda t, \tag{6}$$

in which $(\boldsymbol{z}, \boldsymbol{q})$ satisfy assumptions analogous to $(\boldsymbol{Z}, \boldsymbol{Q})$ [2, 13].

The fluid limit model $\mathcal{L}$ is said to be *stable* if there exists $\Omega_0 \subset \Omega$ satisfying $\mathbb{P}\{\Omega_0\} = 1$, and $T_0 > 0$ such that $q(t) = 0$ whenever $t \geq T_0$, $\omega \in \Omega_0$, $\boldsymbol{q} \in \mathcal{L}(\omega)$ and $\|q(0)\| = 1$.

Naturally, the fluid limit model as well as the conditions for stability may depend on the scheduling policy. In many cases (such as the G/G/1 queue) the set of fluid limits $\mathcal{L}_x$ is a deterministic singleton for each x.

The following is the key motivating result for our paper.

Theorem 2.2. *Consider the network model* (2) *controlled using a stationary non-idling policy. Suppose that $\mu_i < \infty$ at each exit buffer, and that the fluid limit model is stable. Then,*

(i) $\boldsymbol{Q}$ is aperiodic and positive Harris recurrent*: There is a unique invariant measure π such that the distributions converge in total variation norm for each initial condition $x \in \mathbb{Z}^N_+$,*

$$\lim_{t\to\infty} \Big(\sup_{y\in\mathbb{Z}^N_+} |\mathbb{P}_x(Q(t) = y) - \pi(y)| \Big) = 0.$$

(ii) The invariant measure has polynomial moments*: For each $p \geq 1$,*

$$\sum_{x\in\mathbb{Z}^N_+} \pi(x)\|x\|^p < \infty. \tag{7}$$

Theorem 2.2 asserts that the model (2) is positive Harris recurrent with polynomial moments of order p for *every integer p* when the interarrival and service times are exponentially distributed. This suggests that π will have an exponential tail, so that $\mathbb{E}_\pi[e^{\theta'\|Q(0)\|}] < \infty$ for some $\theta' > 0$. We now show that this conjecture does not hold true.

Theorem 4.1 of [3] considers general models with renewal inputs, and also establishes rates of convergence to stationarity and other ergodic theorems for the model. In Theorem 2.2 we have extracted the part that is most relevant to the counterexample described next.

Theorem 2.3. *Consider the Rybko-Stolyar model described by the Markovian model* (2) *as follows,*

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t), \quad i = 1, 3$$
$$Q_i(t) = Q_i(0) + D_{i-1}(t) - D_i(t), \quad i = 2, 4. \tag{8}$$

There exist network parameters and a stationary policy satisfying,

1. *The interarrival and service times are mutually independent with exponential distribution.*

2. *The fluid limit model is stable (hence there exists a unique invariant measure π.)*

3. *The invariant measure π satisfies*

$$\mathbb{E}_\pi[\Psi(\|Q(0)\|)] = \infty, \tag{9}$$

where $\Psi(s) = s^{\log s}$ for $s > 0$, and $\Psi(0) = 0$. In particular, the invariant measure π does not have an exponential tail and the Markov process is not exponentially ergodic.

Theorem 2.3 establishes that the result (7) of Theorem 2.2 is nearly tight, modulo the log term in the exponent.

The proof of Theorem 2.3 is technical. It is simplified substantially through our adoption of a relaxation in which the service rates at buffers 1 and 3 are infinite. The resulting process violates the assumptions of Dai and Meyn [3], but we show in the following section that these results carry over to this more general setting to yield Theorem 2.2.

3 Stability of Markovian networks

In this section we establish some general properties of the model (2). In this section only we consider the embedded chain obtained via uniformization. It is known that geometric ergodicity[1] of this chain is equivalent to exponential ergodicity of the process [5].

Uniformization must be applied with care when service rates can be infinite. We define sampling times $\{\tau_n, n \geq 0\}$ corresponding to jumps of a Poisson process derived from the arrival-service process $\{\boldsymbol{A}, \boldsymbol{S}\}$. This is commonly interpreted as sampling at arrival epochs and (real or virtual) service completions. However, in sampling we restrict to non-zero length service completions to avoid sampling twice at the same instant!

Right-continuity implies that $Q(\tau_n) = Q(\tau_n^+)$ for each n.

[1]The definition of g. ergodicity is precisely (3) in discrete time.

It is in fact simplest to ignore most of the network structure, and consider a general Markov chain denoted $\boldsymbol{X}$ on $\mathbb{Z}_+^N$ satisfying a version of the so-called skip-free property [16] along with a standard irreducibility condition.

The skip-free condition is defined with respect to the L_1 norm. Under the assumption that the service rate at any exit buffer is finite it follows that the increments $\big|\|X(t+1)\| - \|X(t)\|\big|$ are bounded in the network model (2) when $\{X(t) := Q(\tau_t) : t \in \mathbb{Z}_+\}$. Note that the increments of the norm $\|X(t+1) - X(t)\|$ are not bounded in a queueing model that allows instantaneous transfer of buffer contents.

The following result establishes Proposition 2.1.

Theorem 3.1. *Suppose that* $\boldsymbol{X}$ *is a Markov chain on* $\mathbb{Z}_+^N$ *satisfying the following*

(i) *It is skip-free in the sense that for some constant* b_0 *and every initial condition,*

$$-b_0 \le \|X(t+1)\| - \|X(t)\| \le b_0, \qquad t \ge 0. \tag{10}$$

(ii) *The chain is* 0*-irreducible, in the sense that*

$$\sum_t P^t(x, 0) > 0 \text{ for each } x. \tag{11}$$

(iii) $\boldsymbol{X}$ *is geometrically ergodic*

Then π *has an exponential tail.*

Proof. The proof proceeds in two steps.

Denote the moment generating function by $h_\theta(x) = \mathbb{E}_x[e^{\theta\tau_0}]$, $\theta > 0$, $x \in \mathbb{Z}_+^n$, where τ_0 is the first hitting time to the origin. Theorem 15.2.4 and Theorem 16.3.2 of [16] imply that $\pi(h_\theta) < \infty$ for sufficiently small $\theta > 0$.

The second step is to compare h_θ with the function $g_\theta(x) = e^{\theta\|x\|}$, $\theta > 0$, $x \in \mathbb{Z}_+^N$. The skip-free assumption implies the bound,

$$\big|\|X(t)\| - \|X(0)\|\big| \le b_0 t, \qquad t \ge 0, \tag{12}$$

so that $\tau_0 \ge b_0^{-1}\|x\|$ with probability one when $X(0) = x$. Hence $g_{\theta/b_0}(x) \le h_\theta(x)$ for all x. We conclude that,

$$\mathbb{E}_\pi\left[e^{(\theta/b_0)\|X(0)\|}\right] = \sum \pi(x) g_{\theta/b_0}(x) \le \sum \pi(x) h_\theta(x) < \infty.$$

□

The definition of the fluid limit model is defined for any Markov chain exactly as in the network model via (5).

Theorem 3.2. *Suppose that* $\boldsymbol{X}$ *is a Markov chain on* $\mathbb{Z}_+^N$ *satisfying (i) and (ii) of Theorem 3.1. Suppose moreover that the fluid limit model is stable, in the sense that for some* $T_0 > 0$ *and a set* $\Omega_0 \subset \Omega$ *satisfying* $\mathbb{P}\{\Omega_0\} = 1$,

$$\lim_{\kappa\to\infty} \frac{1}{\kappa}\|X(\kappa t; \kappa x^\kappa)\| = 0, \qquad \|x\| \le 1,\ t \ge T_0,\ \omega \in \Omega_0.$$

Then $\boldsymbol{X}$ *is positive Harris recurrent, and its unique invariant measure* π *satisfies* $\sum \pi(x)\|x\|^p < \infty$ *for each* $p \ge 1$.

Proof. In Proposition 3.3 we establish a Lyapunov drift condition of the form: For each $p \ge 1$ we can find a function V and positive constants b and ϵ such that,

$$PV(x) = \mathbb{E}[V(X(t+1)) \mid X(t) = x] \le V(x) - \epsilon V^{1-\delta}(x) + b\,, \tag{13}$$

where $\delta = (1+p)^{-1}$. Moreoever, the function V is equivalent to $\|x\|^{p+1}$:

$$0 < \liminf_{r\to\infty}\Big(\inf_{\|x\|=r} \frac{V(x)}{r^{p+1}}\Big) \le \limsup_{r\to\infty}\Big(\sup_{\|x\|=r} \frac{V(x)}{r^{p+1}}\Big) < \infty. \tag{14}$$

It follows from the Comparison Theorem of [16] that the steady-state mean of $V^{1-\delta}$ is finite, with the explicit bound $\pi(V^{1-\delta}) \le b/\epsilon$. This implies that the pth moment of $\boldsymbol{X}$ is finite since $1-\delta = p/(p+1)$. ☐

The drift criterion (13) was introduced in the analysis of general state-space Markov chains in [4]. Under this bound polynomial rates of convergence are obtained on the rate of convergence to steady state. Using different methods, polynomial bounds on the steady-state buffer lengths and polynomial rates of convergence were obtained in [3] for stochastic networks based on a general result of [21]. The proof is simplified considerably in this countable state space setting.

To establish (13) we take $V = V_p$ with,

$$V_p(x) = \mathbb{E}\Big[\sum_{t=0}^{\|x\|T} \|X(t)\|^p\Big], \qquad x \in \mathbb{Z}_+^N, \tag{15}$$

where $T \ge 1$ is a sufficiently large fixed integer. The growth bounds (14) are suggested by the approximation,

$$\frac{1}{\kappa^{p+1}} V_p(\kappa x^\kappa) \approx \mathbb{E}\Big[\int_0^{\|x\|T} \|x^\kappa(t; x^\kappa)\|^p\, dt\Big], \qquad x \in \mathbb{R}_+^N,\ \kappa > 0, \tag{16}$$

with $\boldsymbol{x}^\kappa$ defined as in (5) via $x^\kappa(t; x^\kappa) := \kappa^{-1} X(\kappa t; \kappa x^\kappa)$.

Proposition 3.3. *The following hold under the assumptions of Theorem 3.2: for each* $p = 1, 2, \ldots$ *the function* V *defined in* (15) *satisfies the drift condition* (13) *and the bounds* (14) *with* $\delta = (1+p)^{-1}$.

To prove the proposition we first note that stability of the fluid model implies convergence in an L_p sense. The proof of the almost sure limit in Proposition 3.4 uses equicontinuity of $\{x^\kappa(t;x^\kappa) : t \geq 0\}$, and the L_p limit is obtained using the Dominated Converence Theorem since $\boldsymbol{L}$ is a bounded sequence. Recall that T_0 and Ω_0 are introduced in the definition of stability for the fluid limit model.

Proposition 3.4. *Suppose that the fluid model is stable. Then it is uniformly stable in the following two senses,*

(i) Almost surely: *For $T \geq T_0$ and $\omega \in \Omega_0$,*

$$\lim_{\kappa\to\infty} \sup_{\|x\|=1} \|x^\kappa(T;x^\kappa,\omega)\| = 0.$$

(ii) In the L_p sense: *For $T \geq T_0$,*

$$\lim_{\kappa\to\infty} \sup_{\|x\|=1} \mathbb{E}[\|x^\kappa(T;x^\kappa)\|^p] = 0. \tag{17}$$

□

Proof of Proposition 3.3. Before proceeding it is helpful to review the Markov property: Suppose that $\mathcal{C} = \mathcal{C}(X(0), X(1), \dots)$ is any random variable with finite mean. We always have,

$$\mathbb{E}_{X(n)}[\mathcal{C}] = \mathbb{E}[\vartheta^n\mathcal{C} \mid X(0), \dots, X(n)] = \mathbb{E}[\vartheta^n\mathcal{C} \mid X(n)],$$

where $\mathcal{F}_n := \sigma\{X(0), \dots, X(n)\}$, $n \geq 0$, and $\vartheta^n\mathcal{C}$ denotes the random variable,

$$\vartheta^n\mathcal{C} = \mathcal{C}(X(n), X(n+1), \dots)$$

We apply the Markov property with $n = 1$ and $\mathcal{C} = \sum_{t=0}^{\|X(0)\|T} \|X(t)\|^p$. In this case we have,

$$\vartheta^1\mathcal{C} = \sum_{t=1}^{\|X(1)\|T} \|X(t)\|^p$$

so that the Markov property gives,

$$V_p(X(1)) = \mathbb{E}_{X(1)}\big[\mathcal{C}\big] = \mathbb{E}\Big[\sum_{t=1}^{\|X(1)\|T} \|X(t)\|^p \mid \mathcal{F}_1\Big].$$

Applying the transition matrix to V_p gives $PV_p(x) = \mathbb{E}_x[V_p(X(1))]$, so that

$$PV_p(x) = \mathbb{E}_x\Big[\sum_{t=1}^{\|X(1)\|T} \|X(t)\|^p\Big], \qquad x \in \mathbb{Z}_+^N.$$

The sum within the expectation on the right hand side is almost the same as used in the definition of V_p. However, instead of summing from 0 to $T\|X(0)\|$, we are summing from 1 to $T\|X(1)\|$. Consequently, writing $y = X(1)$, we have the expression,

$$PV_p(x) = V_p(x) - \|x\|^p + \mathbb{E}_x\Big[\sum_{t=\|x\|T+1}^{\|y\|T} \|X(t)\|^p\Big], \qquad x \in \mathbb{Z}_+^N, \tag{18}$$

where the sum is interpreted as negative when $\|x\| \geq \|y\|$. Under the assumption that $\boldsymbol{L}$ is a bounded sequence we obtain the bound $PV_p \leq V_p - \|x\|^p + b_p$, where b_p is the supremum over x on the expectation on the right hand side of (18). We now argue that $b_p < \infty$. Recall that under assumption (10) the increments of $\boldsymbol{X}$ are bounded, $|\|x\|T - \|X(1)\|T| \leq b_0 T$. This combined with (17) implies that b_p is indeed finite.

To complete the proof we now establish (14). For this we apply (12), which implies that V_p satisfies the pair of bounds,

$$\sum_{t=0}^{\|x\|T} \|(x - b_0 t)_+\|^p \leq V_p(x) \leq \sum_{t=0}^{\|x\|T} \|x + b_0 t\|^p, \qquad x \in \mathbb{Z}_+^N.$$

This implies (14). □

4 Scheduling policy

The remainder of the paper is devoted to establishing the main result, Theorem 2.3. Throughout the remainder of the paper we restrict attention to the KSRS model (8) in continuous time. The network is assumed symmetric with $\lambda_1 = \lambda_3 = 1$, $\mu_1 = \mu_3 = \infty$, and $\mu_2 = \mu_4$ finite. The traffic intensity of each server is denoted,

$$\begin{aligned}\rho_1 &:= \lambda_1\mu_1^{-1} + \lambda_3\mu_4^{-1} = \mu_4^{-1} \\ \rho_2 &:= \lambda_1\mu_2^{-1} + \lambda_3\mu_2^{-1} = \mu_2^{-1}\end{aligned}$$

To prove Theorem 2.3 we construct a particular non-idling stationary policy.

The proofs of Lemmas 4.1 and 4.2 are contained in the Appendix. Recall that the function Ψ is defined in Theorem 2.3.

Lemma 4.1. *For each constant $c > 1$, the function $\Psi(cs)/\Psi(s)$ is a strictly increasing function. Moreover,*

$$\sum_{1\leq m<\infty} m^2 \Big[\frac{\Psi(m^\eta)}{\Psi(cm^\eta)}\Big]^{\frac{1}{4}} < \infty,$$

for any η satisfying,

$$\eta > \sqrt{\frac{12}{\log(c)}} \tag{19}$$

Throughout the remainder of the paper we fix μ_2, μ_4 so that the conditions specified in Lemma 4.2 are satisfied.

Lemma 4.2. *Define* $\beta_1 = \frac{1}{4}\gamma_{24}$ *and* $\beta_2 = 4\gamma_{24}$ *with,*

$$\gamma_2 = \frac{\rho_2}{1-\rho_2}, \quad \gamma_4 = \frac{\rho_1}{1-\rho_1}, \quad \gamma_{24} = \gamma_2\gamma_4, \quad \gamma = \gamma_4 + \gamma_{24}.$$

The parameters μ_2, μ_4 *can be chosen so that* $\rho_i \in (1/2, 1)$ *for* $i = 1, 2$, $\beta_2 > \beta_1 > 1$, *and the condition (19) is satisfied with* $c = \beta_1$ *and* $\eta = \log(\beta_1)/\log(\beta_2)$.

We define for $s, n \geq 1$,

$$\begin{aligned} \Psi^*(s) &= (\Psi(\beta_1 s^\eta)/\Psi(s^\eta))^{\frac{1}{4}} \\ \psi(n) &= \Psi^*(n)/\Psi^*(n+1). \end{aligned} \tag{20}$$

By Lemma 4.1, Ψ^* is a strictly increasing function, so $\psi(n) \in (0,1)$. The scheduling decisions are parametrized by $\psi = \{\psi(n)\}$ and are made only at the sampling instances $\{\tau_n, n \geq 0\}$ introduced in Section 3.

The scheduling decisions for server 1 at time τ_n are defined as follows:

(i) If only one of the two buffers at server 1 contain jobs (buffer 1 or buffer 4), the server works on the first job in the non-empty buffer. That is, the policy is *non-idling.*

(ii) If both buffer 1 and buffer 4 are non-empty, and buffer 2 is also non-empty, then buffer 1 receives strict priority at server 1. Since $\mu_1 = \infty$, all of the $Q_1(\tau_n)$ jobs are instantly sent to buffer 2.

(iii) If both buffer 1 and buffer 4 are non-empty, and buffer 2 is empty then the scheduling decision depends on whether τ_n corresponds to an arrival into buffer 1 or not. If it does not correspond to an arrival into buffer 1, then server 1 works on the first job in buffer 4.

(iv) The policy is randomized if both buffer 1 and buffer 4 are non-empty, buffer 2 is empty, and τ_n corresponds to an arrival into buffer 1. In this case, with probability $\psi(m)$ the server works on the first job in buffer 4, and with probability $1 - \psi(m)$ it works on the jobs in buffer 1. This choice is made independently from any other randomness in the network. In the second case, since the service rate is infinite, all of the jobs in buffer 1 are instantly sent to buffer 2.

The scheduling decisions in server 2 are defined analogously.

With a slight abuse of notation we denote this scheduling policy by ψ. We denote by $\mathcal{Q} = (Q(0), \mu_2, \mu_4, \psi)$ the queueing network together with the scheduling policy ψ and the initial state $Q(0)$.

The conclusions of Proposition 4.3 are immediate from the discussion in Section 3.

Proposition 4.3. *Under the scheduling policy ψ the process $\boldsymbol{Q}$ and the embedded process $\{X(n) := Q(\tau_n)\}$ are Markovian. The chain $\boldsymbol{X}$ satisfies the skip-free property* (10) *as well as the irreducibility condition* (11)*. If $\boldsymbol{X}$ possesses an invariant measure π then it is necessarily unique, and satisfies $\pi(x^*) > 0$ with,*

$$x^* := (0,0,0,1)^{\mathrm{T}}. \tag{21}$$

The following result is used to translate properties of invariant measures to the process level properties in the network.

Proposition 4.4. *Suppose the queueing network $\mathcal{Q}$ is such that an associated invariant measure π exists, and $\pi(\Psi)$ is finite. Then for every constant $\delta > 0$.*

$$\limsup_{s\to\infty} \Psi(\delta s)\mathbb{P}_{x^*}(\frac{\|Q(s)\|}{s} > \delta) < \infty. \tag{22}$$

Moreover, for every constant $\delta > 0$

$$\limsup_{s\to\infty} \Psi(.5\delta s)\mathbb{P}_{x^*}(\sup_{t\geq s} \frac{\|Q(t)\|}{t} > \delta) < \infty. \tag{23}$$

Proof. By Proposition 4.3, we have $\pi(x^*) > 0$. Applying Markov's inequality gives,

$$\begin{aligned}
\Psi(\delta s)\mathbb{P}_{x^*}(\|Q(s)\| > \delta s) &= \Psi(\delta s)\mathbb{P}_{x^*}(\Psi(\|Q(s)\|) > \Psi(\delta s)) \\
&\leq \mathbb{E}_{x^*}[\Psi(\|Q(s)\|)] \\
&\overset{(*)}{\leq} \pi^{-1}(x^*) \sum_y \mathbb{E}_y[\Psi(\|Q(s)\|)]\pi(y) \\
&= \pi^{-1}(x^*)\mathbb{E}_\pi[\Psi(\|Q(0)\|)],
\end{aligned}$$

where in $(*)$ we simply use $1 = \pi^{-1}(x^*)\pi(x^*) \leq \pi^{-1}(x^*)\sum_y \pi(y)$. Then (22) follows immediately.

We now establish (23). Fix $\delta > 0$ and let $\hat{\delta} = \delta/4$. Consider any $s > 0$ and let $s_k = (1+\hat{\delta})^k s, k \geq 0$. We use the following obvious identity,

$$\sup_{s_k \leq t \leq s_{k+1}} \frac{\|Q(t)\|}{t} \leq \frac{\|Q(s_k)\|}{s_k} + \frac{A_1(s_k, s_{k+1}) + A_3(s_k, s_{k+1})}{s_k},$$

with $A_i(s_k, s_{k+1})$ denoting the arrivals to buffer i in the interval $(s_k, s_{k+1}]$. Hence,

$$\mathbb{P}(\sup_{t\geq s} \frac{\|Q(t)\|}{t} > \delta) \leq \sum_{k\geq 0} \mathbb{P}(\frac{\|Q(s_k)\|}{s_k} \geq \delta/2) + \sum_{k\geq 0} \mathbb{P}(\frac{A_1(s_k, s_{k+1}) + A_3(s_k, s_{k+1})}{s_k} \geq \delta/2).$$

To complete the proof it suffices to show that

$$\limsup_{s\to\infty} \Psi(.5\delta s) \sum_{k\geq 0} \mathbb{P}\Big(\frac{\|Q(s_k)\|}{s_k} \geq \delta/2\Big) < \infty, \tag{24}$$

$$\limsup_{s\to\infty} \Psi(.5\delta s) \sum_{k\geq 0} \mathbb{P}\Big(\frac{A_1(s_k, s_{k+1}) + A_3(s_k, s_{k+1})}{s_k} \geq \delta/2\Big) < \infty, \tag{25}$$

Applying (22), we have that

$$\begin{aligned}\sum_k \mathbb{P}_{x^*}\Big(\frac{\|Q(s_k)\|}{s_k} \geq \delta/2\Big) &= O\Big(\sum_k [\Psi(.5\delta s_k)]^{-1}\Big) = O\Big(\sum_k [\Psi(.5\delta(1+\hat{\delta})^k s)]^{-1}\Big) \\ &= O\Big(\sum_k \big(.5\delta(1+\hat{\delta})^k s\big)^{-\log(.5\delta(1+\hat{\delta})^k s)}\Big)\end{aligned}$$

Now for all s such that $.5\delta s > 1$ we have

$$\begin{aligned}\sum_k \frac{1}{(.5\delta(1+\hat{\delta})^k s)^{\log(.5\delta(1+\hat{\delta})^k s)}} &\leq \sum_k \frac{1}{(.5\delta s)^{\log(.5\delta(1+\hat{\delta})^k s)}} \\ &= \sum_k \frac{1}{(.5\delta s)^{\log(.5\delta s) + k\log(1+\hat{\delta})}} \\ &= (.5\delta s)^{-\log(.5\delta s)} \frac{1}{1-(.5\delta s)^{-\log(1+\hat{\delta})}} \\ &= \Psi^{-1}(.5\delta s)\frac{1}{1-(.5\delta s)^{-\log(1+\hat{\delta})}}.\end{aligned}$$

Since the limit of the second component of the product above is equal to unity, we obtain that (24) holds.

We now establish (25). Using the large deviations bound given by Lemma A.1 in the Appendix, we have

$$\begin{aligned}\mathbb{P}(A_1(s_k, s_{k+1}) \geq (\delta/2)s_k) &= \mathbb{P}(A_1(s_k, s_{k+1}) - (s_{k+1} - s_k) \geq (\delta/2)s_k - (s_{k+1} - s_k)) \\ &\leq \mathbb{P}(A_1(s_k, s_{k+1}) - (s_{k+1} - s_k) \geq (\frac{\delta}{2\hat{\delta}} - 1)(s_{k+1} - s_k)) \\ &= \mathbb{P}(A_1(s_k, s_{k+1}) - (s_{k+1} - s_k) \geq (s_{k+1} - s_k)) \\ &\leq \exp(-\Omega(s_{k+1} - s_k)) \\ &\leq \exp(-\Omega(\hat{\delta}(1+\hat{\delta})^k s))\end{aligned}$$

We also have

$$\sum_{k\geq 0} \exp(-\Omega(\hat{\delta}(1+\hat{\delta})^k s)) \leq \exp(-\Omega(s)).$$

Since Ψ grows slower than exponentially we conclude that (25) holds. $\square$

5 Stability of the fluid limit model

The goal of this section is to establish stability of the fluid limit model obtained from the scheduling policy ψ. Specifically, we establish the following result which is Part 2 of Theorem 2.3.

Theorem 5.1. *There exists a constant* $c_0 > 0$ *such that for every* $\epsilon > 0$

$$\limsup_{\|x\|\to\infty} \mathbb{P}_x\Big(\frac{\|Q(c_0\|x\|)\|}{\|x\|} > \epsilon\Big) = 0.$$

Proof. The result is obtained by combining Proposition 5.3 with Lemma 5.4 that follow. □

The proofs in this section rely on regeneration arguments, based on the stopping times defining the first emptying times for the four buffers in the network,

(26)

$$T_i = \inf\{t \geq 0 : Q_i(t) = 0\} \quad \bar{T}_i = \inf\{t \geq 0 : Q_j(t) = 0 \text{ for all } \ j \neq i,\, i = 1, \ldots, 4\}.$$

The following feature of the model is crucial to obtain regenerative behavior: Although the policy is not a priority policy, its behavior at buffers 4 and 2 is similar. For example, since the policy is non-idling and the service rate at buffer 1 is infinite we can conclude that buffer 4 receives full capacity while this buffer is non-empty.

Lemma 5.2 follows from these observations, and the specification of the policy that gives priority to jobs in buffer 3 while buffer 4 is non-empty.

Lemma 5.2. *Under the policy* ψ *with* $Q_4(0) \geq 1$*, the contents of buffer 4 evolves as an M/M/1 queueing system with arrival rate* 1 *and service rate* μ_4*, until the first time* Q_4 *becomes zero.*

Proposition 5.3 concerns the special case in which all the jobs are initially in buffer 4.

Proposition 5.3. *Let* $x^n = (0,0,0,n) = nx^*$ *and* $\tau = 1/(\mu_4 - 1)$*. For every* $\epsilon > 0$

$$\lim_n \mathbb{P}_{x^n}\Big(\|Q(\tau n)\| \leq \epsilon n\Big) = 1.$$

Proof. We begin with an application of Lemma 5.2 to conclude that $Q_4(t)$ corresponds with an M/M/1 queueing system up to time T_4. Applying Lemma A.2 we obtain for every $\epsilon > 0$

$$\lim_n \mathbb{P}_{x^n}\Big(|T_4 - \frac{n}{\mu_4 - 1}| \leq \epsilon n\Big) = 1. \tag{27}$$

We now analyze the queue-length processes in buffers 1 and 2 during $[0, T_4)$. For each $k \geq 1$, with probability $(1 - \psi(k)) \prod_{1 \leq i \leq k-1} \psi(i)$, priority is given to jobs in buffer 4 up until the time of the k-th arrival. At this time the k jobs at buffer 1 are immediately transfered to buffer 2. While buffer 2 is non-empty, full priority is given to jobs in buffer 1 over jobs in buffer 4, so that any new arrivals are sent to buffer 2 instantaneously. At the first instance $\tau_2(1) < T_4$ that the buffer 2 empties (assuming one exists), this process repeats.

Hence the policy ψ induces the following behavior: there is an alternating sequence $R_1, L_1, R_2, L_2, \ldots$, where R_l corresponds to time-intervals before T_4 when buffer 2 is empty, and L_l corresponds to time-intervals before T_4 when buffer 2 is non-empty. The sequences $R_l, l \geq 1$ and $L_l \geq l$ are i.i.d., and the length of R_l is independent from $R_{l'}, L_{l'}, l' \leq l-1$.

Let $\tau_l = \sum_{j \leq l}(R_j + L_j)$. The queue length in buffer 1 at the end of the time period corresponding to R_l (that is at time $\tau_{l-1} + R_l$) is equal to k with probability $(1-\psi(k)) \prod_{1 \leq i \leq k-1} \psi(i)$. Let X_l represent this queue length $Q_1(t)$ at time $t = \tau_{l-1} + R_l$. We have

$$
\begin{aligned}
\mathbb{E}[X_l^2] &= \sum_{k \geq 1} k^2 (1 - \psi(k)) \prod_{1 \leq i \leq k-1} \psi(i) \\
&< \sum_{k \geq 1} k^2 \prod_{1 \leq i \leq k-1} \psi(i) \\
&= \sum_{k \geq 1} \frac{k^2 \Psi^*(1)}{\Psi^*(k)}
\end{aligned}
$$

Applying the second part of Lemma 4.1, this sum is finite, namely X_l has a finite second moment. Then R_l has finite second moment as well since, conditioned on $X_l = k$, it is a sum of k i.i.d. random variables with $\text{Exp}(\lambda)$ distribution. Conditioning on $X_l = k$, the length of L_l represents the time to empty an M/M/1 queueing system with k initial jobs, arrival rate 1, and service rate μ_2. Hence $\mathbb{E}[L_l^2 | X_l = k] = O(k^2)$ by Lemma A.2. Since X_l has a finite second moment, so does L_l. We conclude that $\tau_l - \tau_{l-1} = R_l + L_l$ has a finite second moment.

Assume now that $T_4 = \infty$ by placing infinitely many jobs in buffer 4. Given any positive $t > 0$, let $l^*(t)$ be the unique index such that $t \in [\tau_{l^*(t)-1}, \tau_{l^*(t)}]$. Applying Smith's Theorem for regenerative processes (see Theorem 3.7.1. in Resnick [17]), for every $m \geq 0$

$$
\lim_{t \to \infty} \mathbb{P}(Q_1(t) + Q_2(t) \geq m) = \frac{\mathbb{E}[\int_0^{\tau_1} 1\{Q_1(t) + Q_2(t) \geq m\} dt]}{\mathbb{E}[\tau_1]},
$$

where $\tau_1 = R_1 + L_1$. Applying the Cauchy-Schwartz inequality we obtain,

$$\begin{aligned}\mathbb{E}[\int_0^{\tau_1} 1\{Q_1(t)+Q_2(t)\geq m\}dt &\leq \mathbb{E}[1\{\sup_{0\leq t\leq \tau_1} Q_1(t)+Q_2(t)\geq m\}\tau_1] \\ &\leq \sqrt{\mathbb{P}(\sup_{0\leq t\leq\tau_1} Q_1(t)+Q_2(t)\geq m)\mathbb{E}[\tau_1^2]} \\ &\leq \sqrt{\mathbb{P}(A_1(\tau_1)\geq m)\mathbb{E}[\tau_1^2]},\end{aligned}$$

Observe that $A_1(\tau_1) = X_1 + A(R_1, \tau_1)$. Conditioning on $X_1 = k$ and applying Lemma A.2 gives $\mathbb{E}[A(R_1, \tau_1)] = O(k)$, and since the mean of X_1 is finite, $\mathbb{E}[X_1] < \infty$, we obtain $\mathbb{E}[A_1(\tau_1)] < \infty$.

Markov's inequality, gives the bound $\mathbb{P}(A_1(\tau_1) \geq m) = O(1/m)$, and since $\mathbb{E}[\tau_1^2] < \infty$ we conclude,

$$\lim_{t\to\infty} \mathbb{P}_{x^\infty}(Q_1(t) + Q_2(t) \geq m) = O(\frac{1}{m}).$$

Now we recall (27) and use independence of interarrival and service times to conclude that

$$\lim_n \mathbb{P}_{x^n}(Q_1(T_4) + Q_2(T_4) \geq m) = O(\frac{1}{m}).$$

Since $Q_3(T_4) + Q_4(T_4) = 0$, we obtain an apparently weaker result that for every $\epsilon > 0$

$$\lim_n \mathbb{P}_{x^n}\big(\|Q(T_4)\| \leq \epsilon n\big) = 1.$$

It remains to relate $\|Q(T_4)\|$ to $\|Q(\frac{n}{\mu_4-1})\|$. We use the fact that the difference between $\|Q(\frac{n}{\mu_4-1})\|$ and $\|Q(T_4)\|$ is at most the total number of arrivals and departures during the time interval between $\frac{n}{\mu_4-1}$ and T_4. Using the large deviations bound Lemma A.1 applied to arrival processes to buffers 1 and 3 and service processes in buffers 2 and 4, and combining with (27), we obtain

$$\lim_n \mathbb{P}_{x^n}\big(\big\|Q\big(\frac{n}{\mu_4 - 1}\big)\big\| \leq \epsilon n\big) = 1.$$

□

We now assume that x is not of the form $(0, 0, 0, n)$ and complete the proof of Theorem 5.1 by establishing the following lemma.

Recall that $\{\bar{T}_i\}$ are defined in (26), given $Q(0) = x$. We let T denote the minimum,

$$T = \min(\bar{T}_2, \bar{T}_4).$$

Lemma 5.4. *For some constants $c_1, c_2 > 0$,*

$$\liminf_{\|x\|\to\infty} \mathbb{P}_x\big(T \leq c_1\|x\| \text{ and } \|Q(T)\| \leq c_2\|x\|\big) = 1.$$

Proof. To obtain bounds on the probability in the lemma we claim it is enough to consider the special case in which one of the servers has no jobs at time $t = 0$.

Suppose both of the servers are initially non-empty, let $x = Q(0)$, and consider the following cases. If $Q_4(0) = 0$ then this violates the right-continuity assumption since we then have $Q_1(0) > 0$, and all of the jobs in buffer 1 proceed immediately to buffer 2. In the second case $Q_4(0) > 0$, and we can apply Lemma 5.2 to deduce that $T_4(x) = O(\|x\|)$ with probability approaching unity for large $\|x\|$. At time T_4 buffer 3 is empty and all of the jobs in buffer 1 immediately proceed to buffer 2. Once again we arrive at a state $Q(T_4)$ corresponding to an empty server.

In the remainder of the proof we assume that one server is empty at time $t = 0$; Without loss of generality this is server 2, so that $x_2 + x_3 = 0$. Again, if in addition $x_4 = 0$ then only buffer 1 is non-empty, and all the jobs in buffer 1 are sent to buffer 2, giving $T = 0$.

Otherwise, suppose that $x_2 + x_3 = 0$, $x_4 > 0$, and consider the queue length process at buffer 4. Applying Lemma 5.2 we conclude that buffer 4 evolves as an M/M/1 queueing system, and Lemma A.2 implies that the emptying time T_4 for buffer 4 is less than $c_1\|x\|$, with probability approaching unity as $n \to \infty$, for some constant c_1. At $t = T_4$ buffer 3 is empty as well and all the jobs in buffer 1 (if any) instantly proceed to buffer 2. Hence buffers 1,3 and 4 are all empty at time T_4, so that $T = T_4$.

This gives a uniform bound on the probability $\mathbb{P}_x\big(T \le c_1\|x\|\big)$ for initial conditions corresponding to one empty server. To obtain a uniform bound on $\mathbb{P}_x\big(\{T \le c_1\|x\|\} \cap \{\|Q(T)\| \le c_2\|x\|\}\big)$ for some c_2 we apply the large deviations bound of Lemma A.1 to the arrival processes along with the bound $\|Q(t)\| \le \|x\| + \|A(t)\|$. $\square$

6 Lower bounds on the tail probability

The goal of this section is to prove Part 3 of Theorem 2.3. In light of Theorem 3.1 it suffices to establish (9), and then the lack of exponential ergodicity will follow.

Recall the definition of the first emptying time T_4 from (26). We fix a constant $\epsilon > 0$, and define the following events given $Q(0) = x$,

$$\mathcal{E}_4 = 1\Big\{\frac{(1-\epsilon)x_4}{\mu_4 - 1} \le T_4 \le \frac{(1+\epsilon)x_4}{\mu_4 - 1}\Big\},$$
$$\mathcal{E}_1 = 1\Big\{\frac{(1-\epsilon)^2 x_4}{\mu_4 - 1} \le Q_2(T_4) \le \frac{(1+\epsilon)^2 x_4}{\mu_4 - 1} \wedge Q_i(T_4) = 0,\ i \ne 2\Big\}.$$

Like Lemma 5.4, the following result compares an emptying time for the stochastic model with the emptying time for a fluid model.

Lemma 6.1. *Consider a non-zero initial state $x = (x_1, x_2, x_3, x_4)$ satisfying $x_i = 0, i \ne 4$. Then for any $0 < \epsilon < 1/10$*

$$\mathbb{P}_x(\mathcal{E}_4 \cap \mathcal{E}_1) \ge \Theta((\Psi^*(2\gamma_4 x_4))^{-1}). \tag{28}$$

Proof. Consider first a modified scheduling policy $\tilde{\psi}$ that always gives priority to jobs in buffer 4 in server 1 and buffer 3 in server 2. For the policy $\tilde{\psi}$ we let $\{\tilde{T}_i\}$ denote the draining times for the four buffers. For this policy all jobs in buffer 3 are transfered to buffer 4 instantaneously so that buffer 4 operates as an M/M/1 queueing system for all $t \geq 0$. Applying the bound (39) from Lemma A.2, the stopping time $\tilde{T}_4$ satisfies,

$$\mathbb{P}\Big(\frac{(1-\epsilon)x_4}{\mu_4 - 1} \leq \tilde{T}_4 \leq \frac{(1+\epsilon)x_4}{\mu_4 - 1}\Big) \geq 1 - \exp(-\Omega(x_4)). \tag{29}$$

Let $m_1(x)$ denote the number of arrivals to buffer 1 before buffer 4 becomes empty for $\tilde{\psi}$. That is,

$$m_1(x) = A_1(\tilde{T}_4).$$

Using Lemma A.1 applied to the arrival process, combined with (29), we obtain

$$\mathbb{P}\Big(\frac{(1-\epsilon)^2 x_4}{\mu_4 - 1} \leq m_1(x) \leq \frac{(1+\epsilon)^2 x_4}{\mu_4 - 1}\Big) \geq 1 - \exp(-\Omega(x_4)). \tag{30}$$

Now we return to the original scheduling policy ψ. For every m, conditioned on $m_1(x) = m$, the probability that at every arrival instance the priority was given to jobs in buffer 4 is $\prod_{1 \leq i \leq m} \psi(i) = (\Psi^*(m))^{-1}$. If it is indeed the case that at every arrival into buffer 1 the priority was given to buffer 4 for all arrivals up to the $m_1(x)$-th, then $T_4 = \tilde{T}_4$ and $Q_1(T_4) = m_1(x)$. If in addition $m = m_1(x)$ satisfies the bound

$$m \leq \frac{(1+\epsilon)^2 x_4}{\mu_4 - 1} = \frac{(1+\epsilon)^2 \rho_1 x_4}{1 - \rho_1} < 2\gamma_4 x_4,$$

then, by monotonicity of Ψ^*, we have $(\Psi^*(m))^{-1} \geq (\Psi^*(2\gamma_4 x_4))^{-1}$. If during the time interval $[0, T_4]$ server 1 is processing only jobs in buffer 4, then at time T_4 server 2 is empty. Since at time T_4 buffer 4 also becomes empty, all the $Q_1(T_4)$ jobs in buffer 1 instantly arrive into buffer 2 at time T_4, and all the other buffers become empty at T_4. That is $Q_i(T_4) = 0, i \neq 2$ and $Q_2(T_4) = m_1(x)$. These arguments combined with (29) and (30) give,

$$\mathbb{P}_x(\mathcal{E}_4 \cap \mathcal{E}_1) = (\Psi^*(2\gamma_4 x_4))^{-1} - 2\exp(-\Omega(x_4)) = \Theta((\Psi^*(2\gamma_4 x_4))^{-1}),$$

where the last equality follows from subexponential decay of Ψ^*. This completes the proof. □

Define the stopping time,

$$T_{24} = \inf\{t > T_4 : Q_2(t) = 0\}, \tag{31}$$

and for a given $x = Q(0)$ consider the events,

$$\mathcal{E}_{42} = 1\Big\{\frac{(1-\epsilon)^3 x_4}{(\mu_4-1)(\mu_2-1)} \le T_{24} - T_4 \le \frac{(1+\epsilon)^3 x_4}{(\mu_4-1)(\mu_2-1)}\Big\},$$
$$\mathcal{E}_{13} = 1\Big\{\frac{(1-\epsilon)^4 x_4}{(\mu_4-1)(\mu_2-1)} \le Q_4(T_{24}) \le \frac{(1+\epsilon)^4 x_4}{(\mu_4-1)(\mu_2-1)} \wedge Q_i(T_{24}) = 0,\ i \ne 4\Big\}.$$

Lemma 6.2. *Consider a starting state $x = (x_1, x_2, x_3, x_4)$ satisfying $x_i = 0, i \ne 4$. Then for every $\epsilon > 0$*

$$\mathbb{P}_x(\mathcal{E}_{42} \cap \mathcal{E}_{13} \cap \mathcal{E}_4 \cap \mathcal{E}_1) = \Theta((\Psi^*(4\gamma_{24}x_4))^{-2}). \tag{32}$$

Proof. The event $\mathcal{E}_1$ is equivalent to $\{Q(T_4) = (0, z, 0, 0) \text{ for some } z \in \mathcal{S}_4\}$, where $\mathcal{S}_4$ is the set of integers,

$$\mathcal{S}_4 = \{z \in \mathbb{Z}_+ : \frac{(1-\epsilon)^2 x_4}{\mu_4 - 1} \le z \le \frac{(1+\epsilon)^2 x_4}{\mu_4 - 1}\}.$$

Consequently,

$$\begin{aligned}\mathbb{P}\Big(\mathcal{E}_{42} \cap \mathcal{E}_{13} \cap \mathcal{E}_4 \cap \mathcal{E}_1\Big) &= \sum_{z \in \mathcal{S}_4} \mathbb{P}\Big(\mathcal{E}_{42} \cap \mathcal{E}_{13} \cap (Q(T_4) = (0, z, 0, 0)) \cap \mathcal{E}_4\Big) \\ &= \sum_{z \in \mathcal{S}_4} \mathbb{P}\Big(\mathcal{E}_{42} \cap \mathcal{E}_{13} \Big| Q(T_4) = (0, z, 0, 0)\Big)\mathbb{P}\Big(Q(T_4) = (0, z, 0, 0) \cap \mathcal{E}_4\Big),\end{aligned}$$

where in the second equality we use the Markovian property of the scheduling policy ψ. Applying Lemma 6.1 but interchanging buffer 4 with buffer 2 and buffer 3 with buffer 1, we obtain for every z.

$$\mathbb{P}\Big(\mathcal{E}_{42} \cap \mathcal{E}_{13} \Big| Q(T_4) = (0, z, 0, 0)\Big) \ge \Theta((\Psi^*(2\gamma_2 z))^{-1}).$$

For every $z \in \mathcal{S}_4$, given the bound $\epsilon < 1/10$, we have $z \le 2\gamma_4 x_4$. By monotonicity of Ψ^* and the definition $\gamma_{24} = \gamma_2\gamma_4$ we obtain $[\Psi^*(2\gamma_2 z)]^{-1} \ge [\Psi^*(4\gamma_{24}x_4)]^{-1}$.

Finally, we note that,

$$\sum_{z \in \mathcal{S}_4} \mathbb{P}\Big(\{Q(T_4) = (0, z, 0, 0)\} \cap \mathcal{E}_4\Big) = \mathbb{P}\Big(\mathcal{E}_1 \cap \mathcal{E}_4\Big)$$

which is at least $\Theta((\Psi^*(2\gamma_4 x_4))^{-1})$ by Lemma 6.1. Now $\rho_2 > 1/2$ implies $\gamma_2 > 1$ which gives $\gamma_{24} = \gamma_2\gamma_4 > \gamma_4$. Combining these bounds we obtain the desired bound (32). $\square$

Lemma 6.3. *Given a state x such that $x_i = 0, i \ne 4$, there exists a random time T such that,*

$$\mathbb{P}\Big(\frac{\gamma x_4}{2} \le T \le 2\gamma x_4 \ \wedge\ \frac{\gamma_{24}}{2\gamma} \le \frac{Q_4(T)}{T} \le \frac{2\gamma_{24}}{\gamma} \ \wedge\ Q_i(T) = 0, i \ne 4\Big)$$
$$\ge \Theta((\Psi^*(4\gamma_{24}x_4))^{-2}).$$

Proof. The event $\mathcal{E}_4 \cap \mathcal{E}_{42}$ implies

$$((1-\epsilon)\gamma_4 + (1-\epsilon)^3\gamma_{24})x_4 \le T_{24} \le ((1+\epsilon)\gamma_4 + (1+\epsilon)^3\gamma_{24})x_4,$$

the event $\mathcal{E}_{13}$ implies

$$(1-\epsilon)^4\gamma_{24} \le Q_4(T_{24}) \le (1+\epsilon)^4, \gamma_{24}$$

and the bound $\epsilon < 1/10$ implies that $1/2 < \frac{(1-\epsilon)^4}{(1+\epsilon)^4} < \frac{(1+\epsilon)^4}{(1-\epsilon)^4} < 2$. The result is obtained from Lemma 6.2 and using $\gamma := \gamma_4 + \gamma_{24}$. □

We now use Lemma 6.3 is used to obtain the following bound.

Proposition 6.4. *There exist constants* $0 < \alpha < 1,\ \delta > 0$ *such that for every* $n \in \mathbb{Z}_+$

$$\mathbb{P}_{x^*}\Big[\sup_{t \ge \beta_1^n} \frac{\|Q(t)\|}{t} > \delta\Big] \ge \alpha^n \prod_{1 \le m \le n} (\Psi^*(\beta_2^{m+1}))^{-2}. \tag{33}$$

Proof. For every $n \ge 1$ denote by $\mathcal{E}_n$ the event that the event described in Lemma 6.3 occurs n times in succession starting from the state $x = x^*$. For each m let σ_m is the length of the time interval corresponding to the m-th event, and let $S_n = \sum_{m=1}^n \sigma_m$. The random time S_n is only defined on the event $\mathcal{E}_n$.

Using Lemma 6.3, the event $\mathcal{E}_n$ implies that $\frac{\gamma}{2} \le \sigma_1 \le 2\gamma$, and for each $2 \le m \le n$,

$$\begin{aligned} \frac{\gamma}{2} Q_4\Big(\sum_{j \le m-1} \sigma_j\Big) \le \sigma_m \le 2\gamma Q_4\Big(\sum_{j \le m-1} \sigma_j\Big), \\ \frac{\gamma_{24}}{2\gamma}\sigma_m \le Q_4\Big(\sum_{j \le m} \sigma_j\Big) \le \frac{2\gamma_{24}}{\gamma}\sigma_m. \end{aligned} \tag{34}$$

Combining, we obtain

$$\sigma_n \ge \frac{\gamma_{24}}{4}\sigma_{n-1} \ge \frac{\gamma_{24}^2}{4^2}\sigma_{n-2} \ge \cdots \ge \frac{\gamma_{24}^{n-1}}{4^{n-1}}\sigma_1 \ge \frac{\gamma_{24}^{n-1}}{4^{n-1}}\frac{\gamma}{2}, \tag{35}$$

and

$$\sigma_n \le 4\gamma_{24}\sigma_{n-1} \le (4\gamma_{24})^2\sigma_{n-2} \le \cdots \le (4\gamma_{24})^{n-1}\sigma_1 \le (4\gamma_{24})^{n-1}2\gamma, \tag{36}$$

The chain of inequalities (35) implies in particular that

$$S_n \ge \sigma_n \ge \frac{\gamma_{24}^{n-1}}{4^{n-1}}\frac{\gamma}{2}. \tag{37}$$

Since $\beta_1 = \gamma_{24}/4 > 1$ by Lemma 4.2, the same chain of inequalities implies

$$(38) \qquad S_n = \sum_m \sigma_m \le \sigma_n \sum_{m\le n} \frac{\gamma_{24}^{-(n-m)}}{4^{-(n-m)}} \le \sigma_n \sum_{j=1}^{\infty} \frac{\gamma_{24}^{-j}}{4^{-j}} = \frac{\sigma_n}{1-\frac{4}{\gamma_{24}}}.$$

Also the event $\mathcal{E}_n$ implies

$$\frac{\|Q(S_n)\|}{\sigma_n} \ge \frac{\gamma_{24}}{2\gamma},$$

Then from (38) we obtain

$$\frac{\|Q(S_n)\|}{S_n} \ge \frac{\gamma_{24}}{2\gamma}(1-\frac{4}{\gamma_{24}}).$$

Recalling the bound (37), we conclude that the event $\mathcal{E}_n$ implies

$$\sup\{t^{-1}\|Q(t)\| : t \ge \frac{\gamma}{2}(\frac{\gamma_{24}}{4})^{n-1}\} \ge \frac{\|Q(S_n)\|}{S_n} \ge \frac{\gamma_{24}}{2\gamma}(1-\frac{4}{\gamma_{24}}).$$

We have $\beta_1 = \gamma_{24}/4 < \gamma/2$ and conclude that the event $\mathcal{E}_n$ implies

$$\sup_{t\ge\beta_1^n} \frac{\|Q(t)\|}{t} > \delta := \frac{\gamma_{24}}{2\gamma}(1-\frac{4}{\gamma_{24}}).$$

It remains to obtain the required lower bound on $\mathbb{P}_{x^*}[\mathcal{E}_n]$. We first obtain a bound on $\mathbb{P}_{x^*}[\mathcal{E}_m|\mathcal{E}_{m-1}], m = 2, 3, \ldots, n$. For convenience we set $\mathcal{E}_0 = 1\{Q(0) = x^*\}$. For each $1 \le m \le n$ the event $\mathcal{E}_{m-1}$ via (36) and (34) implies

$$Q_4(\sum_{j\le m-1}\sigma_j) \le \frac{2}{\gamma}\sigma_m \le \frac{2}{\gamma}(4\gamma_{24})^{m-1}2\gamma = 4^m\gamma_{24}^{m-1}.$$

Let $\alpha > 0$ be a constant hidden in the $\Theta(\cdot)$ notation in Lemma 6.3. Then, from Lemma 6.3, we obtain

$$\mathbb{P}_{x^*}[\mathcal{E}_m|\mathcal{E}_{m-1}] \ge \alpha(\Psi^*((4\gamma_{24})4^m\gamma_{24}^{m-1}))^{-2} \ge \alpha(\Psi^*((4\gamma_{24})^{m+1}))^{-2}.$$

Since this holds for every $n \ge m \ge 1$ and $\mathcal{E}_1 \supset \mathcal{E}_2 \supset \cdots \supset \mathcal{E}_n$ we obtain

$$\mathbb{P}_{x^*}[\mathcal{E}_n] \ge \alpha^n \prod_{1\le m\le n} (\Psi^*((4\gamma_{24})^{m+1}))^{-2}.$$

This concludes the proof of the proposition by substituting β_2 for $4\gamma_{24}$. □

We are ready to conclude the proof of the final part of Theorem 2.3.

Proof of Part 3 of Theorem 2.3. Using the expression (20) for Ψ^* we have

$$\begin{aligned}\prod_{1\le m\le n}\Big(\Psi^*(\beta_2^{m+1})\Big)^2 &= \prod_{1\le m\le n}\Big(\frac{\Psi(\beta_1\beta_2^{\eta(m+1)})}{\Psi(\beta_2^{\eta(m+1)})}\Big)^{\frac{1}{2}}\\ &= \prod_{1\le m\le n}\Big(\frac{\Psi(\beta_1^{m+2})}{\Psi(\beta_1^{m+1})}\Big)^{\frac{1}{2}}\\ &< \Psi^{\frac{1}{2}}(\beta_1^{n+2}).\end{aligned}$$

¿From Proposition 6.4, we obtain that for all n,

$$(1/\alpha)^n\Psi^{\frac{1}{2}}(\beta_1^{n+2})\mathbb{P}_{x^*}\Big[\sup_{t\ge\beta_1^n}\frac{\|Q(t)\|}{t}>\delta\Big]\ge 1.$$

Using $\beta_1^n=s$ and finding a constant $\eta_2>0$ such that $(1/\alpha)=\beta_1^{\eta_2}$, we obtain

$$\liminf_{s\to\infty} s^{\eta_2}\Psi^{\frac{1}{2}}(\beta_1^2 s)\mathbb{P}_{x^*}\Big[\sup_{t\ge s}\frac{\|Q(t)\|}{t}>\delta\Big]\ge 1.$$

Now

$$s^{\eta_2}\Psi^{\frac{1}{2}}(\beta_1^2 s)=s^{\eta_2}(\beta_1^2 s)^{.5\log(\beta_1^2 s)}=\exp(.5\log^2 s+2\log\beta_1\log s+2\log^2\beta_1+\eta_2\log s).$$

Observe, that the righ-hand side, as a function of s is

$$o(e^{\log^2(.5\delta s)})=o(\Psi(.5\delta s))$$

since the leading term in the first term is $e^{.5\log^2 s}$ and in the second term is $e^{\log^2 s}$. We conclude

$$\liminf_{s\to\infty}\Psi(.5\delta s)\mathbb{P}_{x^*}\Big[\sup_{t\ge s}\frac{\|Q(t)\|}{t}>\delta\Big]=\infty.$$

Applying the second part of Proposition 4.4 we obtain (9). The proof of Theorem 2.3 is complete. □

A Appendix

In this section we provide proofs of some of the elementary results we have used above. Some of the results are well known.

Proof of Lemma 4.1. We have

$$\frac{\Psi(cs)}{\Psi(s)}=\frac{(cs)^{\log(cs)}}{s^{\log s}}=c^{\log(cs)}s^{\log c},$$

which is a strictly increasing function for every $c > 1$. Now suppose $\eta > 0$ is such that (19) is satisfied. Then

$$\sum_{1\le m<\infty} m^2\Big[\frac{\Psi(m^\eta)}{\Psi(cm^\eta)}\Big]^{\frac{1}{4}} = \sum_{1\le m<\infty} m^2\frac{m^{\frac{1}{4}\eta\log m^\eta}}{(cm)^{\frac{1}{4}\eta\log(cm)^\eta}}$$
$$= \sum_{1\le m<\infty}\frac{m^2}{c^{\frac{1}{4}\eta^2\log cm}m^{\frac{1}{4}\eta^2\log c}}.$$

The condition (19) implies $\frac{1}{4}\eta^2\log c - 2 > 1$, which implies the result.

□

Proof of Lemma 4.2. We fix a small $\delta > 0$ and let $\mu_2 = \mu_4 = 1+\delta$. Then $\rho_2 = \rho_1 = 1/(1+\delta), \gamma_2 = \gamma_4 = 1/\delta, \gamma_{24} = 1/\delta^2, \beta_2 = 4/\delta^2, \beta_1 = 1/(4\delta^2), \eta = \frac{\log(1/\delta^2)+4}{\log(1/\delta^2)-4}, c = \beta_1 = 1/(4\delta^2)$. Then

$$\eta^2\log c = \Big(\frac{\log(1/\delta^2)+4}{\log(1/\delta^2)-4}\Big)^2\log(\frac{1}{4\delta^2}).$$

As $\delta \to 0, \eta \to 1, c \to \infty$ and the condition (19) is satisfied for sufficiently small $\delta > 0$. All the other constraints are satisfied trivially.

□

We use in this paper very crude large deviations type bounds. Of course in most cases far more refined large deviations estimates are available [19], but those are not required for our purposes.

Lemma A.1. *Let $N(t)$ be a Poisson process with parameter $\nu > 0$. Then for every constant $\epsilon > 0$*

$$\mathbb{P}(|N(t)-\nu t| > \epsilon t) = \exp(-\Omega(t)).$$

Here the constants hidden in Ω include ϵ.

Lemma A.2. *Consider an M/M/1 queueing system with parameters $\lambda, \mu, \rho = \lambda/\mu < 1$. Let $Q(t)$ denote the queue length at time t and let $T = \inf\{t : Q(t) = 0\}$. Then $\mathbb{E}[T|Q(0) = n] = O(n)$ and $\mathbb{E}[T^2|Q(0) = n] = O(n^2)$. Moreover*

$$\mathbb{P}(\frac{T}{n\mu(1-\rho)} \in (1-\epsilon, 1+\epsilon)) = 1-\exp(-\Omega(n)). \tag{39}$$

Proof. These are well known results from queueing theory. One quick way to establish them is to observe that $T = \sum_{1\le j\le n} T_j$, where T_j is the first passage time from j to $j-1$. That is $T_j = \inf\{t : Q(t) = j-1|Q(0) = j\}$. The sequence T_j is i.i.d. with the distribution equal to the distribution of the busy period, which is known to satisfy $\mathbb{E}[\exp(sT_j)] < \infty$ for some $s > 0$, and has the first moment $1/(\mu(1-\rho))$. We immediately obtain $\mathbb{E}[T|Q(0) = n] = n/(\mu(1-\rho))$ and $\mathbb{E}[T^2|Q(0) = n] = O(n^2)$ (actual value is easy to compute but is not required for our purposes). Applying large deviations bound [19] to the i.i.d. sequence T_j we obtain (39). □

References

[1] H. Chen and D. Yao, *Fundamentals of queueing networks: Performance, asymptotics and optimization*, Springer-Verlag, 2001.

[2] J. G. Dai, *On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid models*, Ann. Appl. Probab. **5** (1995), 49–77.

[3] J. G. Dai and S. P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Transcation on Automatic Controls **40** (1995), 1889–1904.

[4] R. Douc, G. Fort, E. Moulines, and P. Soulier, *Practical drift conditions for subgeometric rates of convergence*, Adv. Appl. Probab. **14** (2004), no. 3, 1353–1377.

[5] D. Down, S. P. Meyn, and R. L. Tweedie. Exponential and uniform ergodicity of Markov processes. *Ann. Probab.*, 23(4):1671–1691, 1995.

[6] N. G. Duffield and Neil O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cambridge Philos. Soc.*, 118(2):363–374, 1995.

[7] D. Gamarnik and A. Zeevi, *Validity of heavy traffic steady-state approximations in open queueing networks*, To appear in Ann. Appl. Prob.

[8] A. Ganesh, N. O'Connell, and D. Wischik. *Big queues*, volume 1838 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2004.

[9] Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Probab.*, 31A:131–156, 1994. Studies in applied probability.

[10] P. R. Kumar and T. I. Seidman. Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automat. Control*, AC-35(3):289–298, March 1990.

[11] P. R. Kumar and S. P. Meyn, *Duality and linear programs for stability and performance analysis queueing networks and scheduling policies*, IEEE Trans. Automat. Control **41** (1996), no. 1, 4–17.

[12] V. A. Malyšev and M. V. Men′šikov, *Ergodicity, continuity and analyticity of countable Markov chains*, Trudy Moskov. Mat. Obshch. **39** (1979), 3–48, 235, *Trans. Moscow Math. Soc.*, pp. 1-48, 1981.

[13] S. P. Meyn. Control techniques for complex networks. To appear, Cambridge University Press, 2007.

[14] S. P. Meyn, *Sequencing and routing in multiclass queueing networks. Part I: Feedback regulation*, SIAM J. Control Optim. **40** (2001), no. 3, 741–776.

[15] S. P. Meyn, *Workload models for stochastic networks: Value functions and performance evaluation*, IEEE Trans. Automat. Control **50** (2005), no. 8, 1106– 1122.

[16] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer-Verlag, 1993.

[17] S. Resnick, *Adventures in stochastic processes*, Birkhuser Boston, Inc., 1992.

[18] A. Rybko and A. Stolyar, *On the ergodicity of stochastic processes describing open queueing networks*, Problemi Peredachi Informatsii **28** (1992), 3–26.

[19] A. Shwartz and A. Weiss, *Large deviations for performance analysis*, Chapman and Hall, 1995.

[20] A. Stolyar, *On the stability of multiclass queueing networks: A relaxed sufficient condition via limiting fluid processe*, Markov Processes and Related Fields (1995), 491–512.

[21] P. Tuominen and R. L. Tweedie, *Subgeometric rates of convergence of f-ergodic Markov chains*, Adv. Appl. Probab. **26** (1994), 775–798.