HAL
archives-ouvertes.fr

Efficiency of simulation in monotone hyper-stable queueing networks

Jonatha Anselmi, Bruno Gaujal

▶ To cite this version:

Jonatha Anselmi, Bruno Gaujal. Efficiency of simulation in monotone hyper-stable queueing networks. Queueing Systems, Springer Verlag, 2013. <hal-00801437>

HAL Id: hal-00801437
https://hal.inria.fr/hal-00801437

Submitted on 16 Mar 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Efficiency of simulation in monotone hyper-stable queueing networks *

Jonatha Anselmi[1] and Bruno Gaujal[2]
[1] Basque Center for Applied Mathematics-BCAM, Al. Mazarredo, 14, Bilbao 48009, Spain
[2] INRIA and LIG, Zirst 51, Av. J. Kuntzmann, MontBonnot Saint-Martin 38330, France
anselmi@bcamath.org, bruno.gaujal@imag.fr

Abstract

We consider Jackson queueing networks with finite buffer constraints (JQN) and analyze the efficiency of sampling from their stationary distribution. In the context of exact sampling, the monotonicity structure of JQNs ensures that such efficiency is of the order of the *coupling time* (or meeting time) of two extremal sample paths. In the context of approximate sampling, it is given by the *mixing time*.

Under a condition on the drift of the stochastic process underlying a JQN, which we call hyper-stability, in our main result we show that the coupling time is polynomial in both the number of queues and buffer sizes. Then, we use this result to show that the mixing time of JQNs behaves similarly up to a given precision threshold. Our proof relies on a recursive formula relating the coupling times of trajectories that start from network states having 'distance one', and it can be used to analyze the coupling and mixing times of other Markovian networks, provided that they are monotone. An illustrative example is shown in the context of JQNs with blocking mechanisms.

1 Introduction

The stationary behavior of Markovian queueing networks (QN) can be computed quite efficiently only under specific assumptions that yield the so called *product-form* property [6, 13]. This property means that the stationary probability distribution of network states can be written, up to a normalizing constant, as the product of a number of simple terms, where each term is associated to a different network node (or queue), and provides a way to compute the stationary behavior of a QN that is much faster than the solution of the global-balance equations of the underlying Markov chain [8]. In several cases, product-form QNs have a restricted modeling power because they often assume that nodes have infinite buffer sizes or that the behavior of a network node does not depend on the state of other nodes; e.g., [12]. Examples of phenomena that do not yield, in general, the product-form property are loss or blocking mechanisms due to finite-buffer constraints or state-dependent routing. On the other hand, the stationary behavior of *non*-product-form QNs is extremely difficult to obtain. While it is possible to obtain it through the solution of a set of linear equations, i.e., the global-balance equations of the underlying Markov chain, the huge size of their state space makes this approach prohibitively expensive from a computational standpoint. For instance, for a QN with 10 nodes where each node has a buffer of 10 units, the number of such equations is not smaller than 11^{10}. In this setting, simulation is a useful approach to obtain robust measures and insights on the stationary behavior.

Existing research in the simulation of Markov chains relies on two approaches: simulation into the future (or Monte Carlo simulation) and simulation from the past (or coupling from the past – CFTP). These are briefly summarized in the following.

Simulation into the future generates a trajectory from *one* state of the chain according to its transition matrix until when it is believed that the proportion of visits on any network state is sufficiently close to its corresponding stationary probability. The *mixing time* [17], i.e., the point in time where the Markov

*Research partially supported by grant MTM2010-17405 (Ministerio de Ciencia e Innovación, Spain) and grant PI2010-2 (Department of Education and Research, Basque Government).

chain is close to its stationary behavior up to a given precision threshold, is a desirable quantity of interest for deciding when to stop the simulation, and it is used to measure the efficiency of simulation into the future. In practice, this type of simulation has the drawbacks of producing excessively-long simulations and approximate estimates of stationary measures.

On the other hand, simulation from the past (or CFTP) generates a trajectory for *each* state of the chain according to its transition matrix and iteratively goes backward in time until when all trajectories have collided in a singleton [20]. The remarkable property of this type of simulation is that such singleton is perfectly distributed according to the stationary distribution of the chain. Then, to infer the stationary behavior of the chain, one can get several independent samples by running several of such simulations and use the law of large numbers. It is important to note that running a trajectory from each state of the chain sounds impractical. However, in *monotone* chains, which is the case of the QNs studied in this paper, it sufficies to generate a trajectory only for each extremal state, and there will be only two of such states in our case. This holds true because monotonicity ensures that the trajectories from non-extremal states are sandwiched between the ones from extremal states.

The major issue behind CFTP is the understanding of the time it takes to produce one sample, which gives a quantitative estimate of the amount of resources that are needed to simulate the Markov chain of interest. In the context of monotone chains, this time is of the order of the *coupling time* (also known as meeting time), i.e., the point in time where the trajectories starting from all possible states collapse into a single one [20].

The mixing time and the coupling time of a Markov chain are related (see Section 4). In this paper, we are interested in bounding these two quantities for a class of important QNs.

1.1 Related work

Since the structure of QNs is arbitrary, exact expressions for the mixing and coupling times are typically difficult to obtain and, therefore, one seeks bounds. In some cases, the mixing time of Markovian QNs with finite state-space can be bounded through general results that involve the second largest eigenvalue modulus of the transition matrix of the chain (see [9, Theorem 3.2]). However, this numerical approach usually hides the understanding of its qualitative dependence with respect to general input parameters.

Qualitative bounds on both the coupling and mixing time of closed Jackson QNs with infinite buffer sizes [8] have been proposed in [10, 16, 15], where the product-form of such networks is the essential tool used for their derivation; see also the references therein for other results concerning product-form QNs. However, the problem becomes much more difficult when buffer sizes are finite because in general the product-form property is lost. In this context, open Jackson QNs with finite buffers (or JQN in the following) have been studied during the last decade. Upper bounds on the mean coupling time of such JQNs are derived in [3], where conditions are given to prove that

i) the coupling time is linear in the total number of queues and exponential in the size of the largest buffer size, and

ii) the coupling time is linear in the sum of the buffer sizes of all queues, provided that the total number of queues is constant.

Linearity in the buffer sizes holds true if the JQN is acyclic, as it has been also shown in [11] in the context of state-dependent routing, or if some other strong conditions on the JQN parameters are satisfied. These conditions, which include restrictions on network topology, will be discussed in Section 3.2. It is conjectured in [3, 21] that linearity in the buffer sizes holds even when the network contains cycles. As numerical experiments reveal, this indicates that the mean coupling time of Markovian QNs can be much smaller than the size of the state space, which is given by the product of the buffer size of each queue, implying that CFTP is efficient for simulating the behavior of JQNs.

1.2 Our contribution

In this paper, we consider JQNs where jobs that try to join a saturated queue are lost. These QNs are intractable in general, as the only exact solution method that is known in the literature relies on the numerical

solution of the global-balance equations of the underlying Markov chain. This amounts to solve $\prod_{i=1}^{M}(C_i+1)$ linear equations (one for each state), where M is the number of queues and C_i is the buffer size of queue i, for $i = 1, \ldots, M$. Matter of fact, the stationary distribution of this class of QNs does not have a product-form; e.g., [3, 11]. To keep the product-form property in networks of queues with finite buffers, one needs special conditions on the topology, the value of the parameters and on the blocking policies used to deal with saturated queues (see for example [5] for a rather exhaustive treatment of such questions). In the general case treated here, no product-form solution exists and the stationary behavior is prohibitively expensive to compute. This motivates our investigation of the efficiency of simulation.

We study the mean coupling time of JQNs in a regime where both the number of queues and buffer sizes vary. In our main result, we give a sufficient condition to prove that the coupling time of JQNs is polynomial in M and the C_i's. This condition lets us deal with networks having arbitrary topology, for which the best bound available in the literature is *exponential* in the C_i's, provided that both M and the C_i's vary [3]. Exploiting classical arguments, we then use this result to prove that the mixing time of JQNs behaves similarly up to a precision threshold.

The starting idea of our proof relies on a new recursive formula on the coupling times of special trajectories having "distance one" at most and follows by stochastic comparison arguments. This allows for a tractable decomposition of the problem that yields our bound. Our approach also provides a new framework for obtaining bounds on the coupling and mixing times of other monotone queueing systems with arbitrary topology, and an example is given in the context of JQNs with blocking [5].

The paper is organized as follows. In Section 2, we give a probabilistic description of JQNs in terms of job movements and an equivalent definition in terms of discrete events. We also briefly review the CFTP method in order to provide the necessary background. In Section 3, we derive our polynomial bound on the coupling time, and, in Section 4, we exploit this bound to derive a bound on the mixing time. Section 5 shows how our approach applies to other queueing network models, and Section 6 draws the conclusions of our work. A preliminary version of this work appeared in [4].

2 Queueing network model

We consider JQNs with M queues. The vector $\mathbf{C} = (C_1, \ldots, C_M)$ denotes the buffer size of each queue. If not otherwise specified, indices i, j, k will implicitly range from 1 to M.

An infinite stream of jobs that follow a Poissonian process with rate λ joins the JQN from an external source. The probability that a job joins queue i, upon arrival to the network, is p_{0i}. Thus, $\sum_i p_{0i} > 0$. In queue i, each job requires some processing for an exponentially distributed amount of time with mean service rate μ_i. The service discipline of each queue i is work-conserving. Upon completion of service at queue i, a job is sent to queue j with probability p_{ij}, and it is accepted if queue j has an available slot (i.e., if it is non-saturated), otherwise it is lost. The probability that a job leaves the network after service at i is p_{i0}, which can be also interpreted as the probability that a job is sent to a queue with buffer of size zero. Since each job eventually leaves the network, $\sum_i p_{i0} > 0$.

The stochastic process $\{(x_1(t), \ldots, x_M(t)) \in \mathbb{Z}^M : 0 \le x_i(t) \le C_i, \forall i\}_{t \ge 0}$, is the continuous-time Markov chain of interest, where state $\mathbf{x}(t) \stackrel{\text{def}}{=} (x_1(t), \ldots, x_M(t))$ denotes the number of jobs in each queue at time t. The space of all the possible states is $\mathcal{S} \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbb{Z}^M : 0 \le x_i \le C_i, \forall i\}$.

2.1 Discrete-event definition of JQN and coupling from the past (CFTP)

The JQN with M queues described above can be seen as a discrete-event system with a single type of events, namely a_{ij}, $(i, j \in \{0, 1, \ldots, M\})$ corresponding to the service of one job in queue i that then joins queue j. The dummy queue 0 corresponds to the outside world. An event of type a_{0j} is an exogenous arrival in queue j and an event of type a_{i0} corresponds to the departure of a job from queue i. If queue i is empty or if queue j is full, then event a_{ij} is disabled. The set of all events is denoted by $\mathcal{A}$.

The rate of event a_{ij} is γ_{ij} and, in view of Remark 1, it is independent of M and $\mathbf{C}$, for any i, j. Using the previous description of a JQN, if $i, j \neq 0$, then $\gamma_{ij} = \mu_i p_{ij}$, $\gamma_{0j} = \lambda p_{0j}$, and $\gamma_{i0} = \mu_i p_{i0}$. The total event rate $\Gamma \stackrel{\text{def}}{=} \sum_{i \ge 0, j \ge 0} \gamma_{ij}$ is finite (we fix $\gamma_{00} = 0$).

We also denote by α_i the expected number of events occurring between two consecutive events involving queue i. By definition, we have

$$\alpha_i = \Gamma/(\gamma_{ii} + \sum_{j=0, j\neq i}^{M}(\gamma_{ji} + \gamma_{ij})), \forall i. \tag{1}$$

The continuous-time Markov chain described above can be transformed into a discrete-time Markov chain $\mathbf{x}_n$ with the same stationary distribution by uniformization by constant Γ. In the following, this discrete chain is assumed to be irreducible and aperiodic. The evolution of the Markov chain $\mathbf{x}_n$ can be written under the form $\mathbf{x}_{n+1} = \phi(\mathbf{x}_n, u_n)$ where u_n is a random variable over the event space that takes value a_{ij} with probability γ_{ij}/Γ, and the transition function $\phi : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is defined as follows:

- If $i, j \neq 0$ then $\phi(\mathbf{x}, a_{ij}) = \mathbf{x} - \mathbf{e}_i + \mathbf{e}_j$ if $0 \leq \mathbf{x} - \mathbf{e}_i + \mathbf{e}_j \leq \mathbf{C}$.
- If $i, j \neq 0$ and $0 \leq \mathbf{x} - \mathbf{e}_i + \mathbf{e}_j$ and $\mathbf{x} - \mathbf{e}_i + \mathbf{e}_j \not\leq \mathbf{C}$ then $\phi(\mathbf{x}, a_{ij}) = \mathbf{x} - \mathbf{e}_i$.
- If $i, j \neq 0$ and $0 \not\leq \mathbf{x} - \mathbf{e}_i$ then $\phi(\mathbf{x}, a_{ij}) = \mathbf{x}$.
- If $i = 0$ then $\phi(\mathbf{x}, a_{0j}) = \mathbf{x} + \mathbf{e}_j$ if $\mathbf{x} + \mathbf{e}_j \leq \mathbf{C}$ and $\phi(\mathbf{x}, a_{0j}) = \mathbf{x}$ otherwise.
- If $j = 0$ then $\phi(\mathbf{x}, a_{i0}) = \mathbf{x} - \mathbf{e}_i$ if $0 \leq \mathbf{x} - \mathbf{e}_i\mathbf{x}$ and $\phi(\mathbf{x}, a_{i0}) = \mathbf{x}$ otherwise.

Let $\phi^{(n)} : \mathcal{S} \times \mathcal{A}^n \to \mathcal{S}$ denote the function whose output is the state of the chain after n iterations starting in state $x \in \mathcal{S}$. That is:

$$\phi^{(n)}(\mathbf{x}, u_{1\to n}) \stackrel{\text{def}}{=} \phi(\ldots\phi(\phi(\mathbf{x}, u_1), u_2), \ldots, u_n).$$

This notation can be extended to sets of states: for $E \subseteq \mathcal{S}$,

$$\phi^{(n)}(E, u_{1\to n}) \stackrel{\text{def}}{=} \left\{\phi^{(n)}(\mathbf{x}, u_{1\to n}), \forall \mathbf{x} \in E\right\}.$$

In the following, $|E|$ denotes the cardinality of set E.

By definition, the function ϕ is monotone for all event $a_{ij} \in \mathcal{A}$. This implies that the trajectories of the Markov chain starting from ordered initial states (using the component-wise ordering) stay ordered: if $\mathbf{x}_0 \leq \mathbf{x}_0^{(1)}$ then $\mathbf{x}_n = \phi^{(n)}(\mathbf{x}_0, u_{1\to n}) \leq \mathbf{x}_n^{(1)} = \phi^{(n)}(\mathbf{x}_0^{(1)}, u_{1\to n})$.

Theorem 1 ([20])**.**

$$\lim_{n\to\infty} \left|\phi^{(n)}(\mathcal{S}, u_{-n+1\to 0})\right| = 1 \textit{ almost surely.}$$

Furthermore, $\phi^{(n)}(\mathcal{S}, u_{-n+1\to 0})$ is steady-state distributed as soon as it is reduced to a singleton, and the mean backward coupling time ($\mathbb{E}\min\{n \in \mathbb{N} | \left|\phi^{(n)}(\mathcal{S}, u_{-n+1\to 0})\right| = 1\}$) is finite.

Theorem 1 has an algorithmic counterpart, namely a coupling from the past (CFTP) algorithm, given as Algorithm 1. In this algorithm, it is implicit that the events u_{-i} are stored and reused along the iterations of the `repeat` cycle. Provided that n is initialized properly, the average time efficiency of Algorithm 1 is $O(\mathbb{E}T)$ where T is the number of iterations in the last `for` loop (called coupling-time in the following); see [20].

The notation used in the remainder of the paper is summarized in Table 1 for quick reference.

3 Bound on coupling time

In this section, we present our main result (Theorem 2) on the efficiency of CFTP applied to the queueing network model described above.

Let

$$\tau(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \min\{n : \phi^{(n)}(\mathbf{x}, u_{1\to n}) = \phi^{(n)}(\mathbf{y}, u_{1\to n})\} \tag{2}$$

be the *coupling time of* $\mathbf{x}$ *and* $\mathbf{y}$, i.e., the random variable of the time where the trajectories starting in states $\mathbf{x} \in \mathcal{S}$ and $\mathbf{y} \in \mathcal{S}$ of a JQN with M queues and buffer sizes $\mathbf{C}$ meet for the first time.

Algorithm 1: Coupling from the past (CFTP)

Data: ϕ , $\{u_{-n}\}_{n\in\mathbb{N}}$
Result: A state sampled from the stationary distribution of the input JQN
begin
  $n = 1$; $m_1 := \mathbf{C}$; $m_2 := 0$;
  repeat
    for $i = n - 1$ **downto** 0 **do**
      $m_1 := \phi(m_1, u_{-i})$;
      $m_2 := \phi(m_2, u_{-i})$;
    $n := 2n$;
  until $m_1 = m_2$;
  return m_1;

The mean coupling time $\mathbb{E}\tau(\mathbf{0}, \mathbf{C})$ is related to the efficiency of simulation from the past. In particular, it is of the order of the average time efficiency of Algorithm 1, as discussed above. The goal of this paper is to bound $\mathbb{E}\tau(\mathbf{0}, \mathbf{C})$ from above with respect to JQNs with arbitrary topology.

In the following, we denote by $\leq_{st}$ the usual stochastic order [19]: for two random variables X_1 and X_2, $X_1 \leq_{st} X_2$ if and only if $\Pr(X_1 \geq x) \leq \Pr(X_2 \geq x)$. The following proposition, known as Strassen's theorem, gives an equivalent characterization of the $\leq_{st}$-order and will be the basic tool used in our proofs.

Proposition 1. *For two random variables X_1 and X_2, $X_1 \leq_{st} X_2$ if and only if there exists a pair of random variables $(\tilde{X}_1, \tilde{X}_2)$ defined on a common probability space such that $X_1 =_{dist} \tilde{X}_1$, $X_2 =_{dist} \tilde{X}_2$, and $Pr(\tilde{X}_1 \leq \tilde{X}_2) = 1$.*

The following proposition is the starting point of our proof technique.

Proposition 2. *For all i and $\mathbf{y} \in \mathcal{S} : y_i > 0$*

$$\tau(\mathbf{0}, \mathbf{y}) \leq_{st} \tau(\mathbf{0}, \mathbf{y} - \mathbf{e}_i) + \max_{k:\mathbf{x}^*+\mathbf{e}_k\in\mathcal{S}} \tau(\mathbf{x}^*, \mathbf{x}^* + \mathbf{e}_k) \tag{3}$$

where $\mathbf{x}^* = \phi^{(n)}(\mathbf{y} - \mathbf{e}_i, u_{1\to n})$ with $n = \tau(\mathbf{0}, \mathbf{y} - \mathbf{e}_i)$.

For $M = 2$ queues, Figure 1 renders a possible illustration of inequality (3) with respect to a sample sequence of events and $\mathbf{y} = \mathbf{C}$, $i = 2$. Using the transition function ϕ, the trajectories starting from $\mathbf{0}$ and $\mathbf{C} - \mathbf{e}_2$ collide at point $\mathbf{x}^*$. By definition, their length is $\tau(\mathbf{0}, \mathbf{C} - \mathbf{e}_2)$. The dashed trajectory starts from $\mathbf{C}$ and uses the same events as the previous ones. At time $\tau(\mathbf{0}, \mathbf{C} - \mathbf{e}_2)$, it has reached point $\mathbf{x}^* + \mathbf{e}_2$. Adding the coupling time of the trajectories starting from points $\mathbf{x}^*$ and $\mathbf{x}^* + \mathbf{e}_2$ provides a bound on $\tau(\mathbf{0}, \mathbf{C})$. The stochastic comparison $\leq_{st}$ follows by a sample path coupling between the trajectories from $\mathbf{C}$ and $\mathbf{C} - \mathbf{e}_i$. In this example, a sequence of events can provided to show that at time $\tau(\mathbf{0}, \mathbf{C} - \mathbf{e}_2)$, the upper trajectory is in point $\mathbf{x}^* + \mathbf{e}_1$. Intuitively, this motivates the max in (3).

Proof. (of Proposition 2). First, we introduce the following lemma.

Lemma 1. *For any possible event $a \in \mathcal{A}$, $k = 1, \ldots, M$, and any state $\mathbf{y}$ such that $\mathbf{y}, \mathbf{y} - \mathbf{e}_k \in \mathcal{S}$,*

$$0 \leq \sum_i (\phi(\mathbf{y}, a)_i - \phi(\mathbf{y} - \mathbf{e}_k, a)_i) \leq 1. \tag{4}$$

Proof. In fact, we have the following cases for every $\mathbf{y} \in \mathcal{S}$ and k such that $\mathbf{y} - \mathbf{e}_k \in \mathcal{S}$:

i) For each queue i, if $a = a_{0i}$ (arrivals) and

  a) if $\mathbf{y} : y_i < C_i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$
  b) if $\mathbf{y} : y_i = C_i$ and $k \neq i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

M	Number of queues
$\mathbf{C}$	$= (C_1, \ldots, C_M)$, vector of the buffer sizes of all queues
λ	Mean arrival rate of jobs
p_{ij}	Probability that a job is forwarded to j upon completion at i,
p_{0i}	Probability that a job joins queue i upon arrival to the network,
p_{i0}	Probability that a job leaves the network after completion at i,
μ_i	Mean job service rate at queue i,
$\mathcal{S}$	$= \{\mathbf{x} \in \mathbb{Z}^M : 0 \leq x_i \leq C_i, \forall i\}$, state space,
$\mathbf{x}, \mathbf{y}$	Generic states,
$\mathcal{A}$	$= \{a_{ij} : 0 \leq i, j \leq M\}$, set of all possible events,
γ_{ij}	$= \mu_i p_{ij}$, $0 \leq i, j \leq M$, rate of the event a_{ij},
α_i	see (1),
u	Random variable over the event space, $\Pr(u = a_{ij}) = \gamma_{ij}/\Gamma$,
$\phi(\mathbf{x}, u)$	$= (\phi(\mathbf{x}, u)_1, \ldots, \phi(\mathbf{x}, u)_M)$, transition function,
$\phi^{(n)}(\mathbf{x}, u_{1\to n})$	n-step transition function,
$\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})$	n-step transition function when $C_i = +\infty$ for all i,
Γ_i	$= \sum_{j=0}^{M} \gamma_{ji}$, minimal upper bound on the mean arrival rate at queue i,
ρ_i	$= \Gamma_i/\mu_i$,
Γ	$= \sum_{i\geq 0, j\geq 0} \gamma_{ij}$, uniformization constant of the Markov chain $(\phi^{(n)}(\mathbf{x}, u_{1\to n}))_{n\in\mathbb{N}}$,
$\tau(\mathbf{x}, \mathbf{y})$	Coupling time of the trajectories starting in states $\mathbf{x}$ and $\mathbf{y}$ (see (2))
$\mathbf{e}_i$	Unit vector in direction i of size M,

Table 1: Summary of the notation used in the paper.

c) if $\mathbf{y} : y_i = C_i$ and $k = i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a)$

ii) For each queue i, if $a = a_{i0}$ (network departures) and

a) if $\mathbf{y} : y_i \geq 0$ and $k \neq i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

b) if $\mathbf{y} : y_i > 1$ and $k = i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

c) if $\mathbf{y} : y_i = 1$ and $k = i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a)$ (here, $y_i = 0$ is not considered because otherwise $\mathbf{y} - \mathbf{e}_k \notin \mathcal{S}$)

iii) For each queue $i > 0$, for each queue $j > 0$, if $a = a_{ij}$ (routings) and

a) if $\mathbf{y} : 1 < y_i \leq C_i$ and $k = i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

b) if $\mathbf{y} : y_i = 1$ and $k = i$ and $y_j < C_j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_j$

c) if $\mathbf{y} : y_i = 1$ and $k = i$ and $y_j = C_j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a)$

d) if $\mathbf{y} : y_j < C_j$ and $k = j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

e) if $\mathbf{y} : y_j = C_j$ and $k = j$ and $y_i \neq 0$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a)$

f) if $\mathbf{y} : y_j = C_j$ and $k = j$ and $y_i = 0$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

g) if $k \neq i$ and $k \neq j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

The above cases cover all the possible situations and prove (4). □

Now, let us compare the trajectories (sample paths) from $\mathbf{y}$, $\mathbf{y} - \mathbf{e}_i$ and $\mathbf{0}$ under the same events. From (4), at the time $n^* = \tau(\mathbf{y} - \mathbf{e}_i, \mathbf{0})$ where the trajectories from $\mathbf{y} - \mathbf{e}_i$ and $\mathbf{0}$ have collided (at state $\mathbf{x}^*$), either $\phi^{(n^*)}(\mathbf{y}, u_{1\to n^*}) = \phi^{(n^*)}(\mathbf{y} - \mathbf{e}_i, u_{1\to n^*})$ or $\phi^{(n^*)}(\mathbf{y}, u_{1\to n^*}) = \phi^{(n^*)}(\mathbf{y} - \mathbf{e}_i, u_{1\to n^*}) + \mathbf{e}_k$ (under the same sequence of events), for some k (see Figure 1 for the latter case where $k = i$ and $\mathbf{y} = \mathbf{C}$). Our upper bound on $\tau(\mathbf{0}, \mathbf{y})$ follows by excluding the possibility of the former case. Taking into account only the latter case, the coupling of the trajectories from $\mathbf{y}$ and $\mathbf{0}$ occurs before $\tau(\mathbf{0}, \mathbf{y} - \mathbf{e}_i)$ plus the additional coupling time of the trajectories starting in $\mathbf{x}^*$ and $\mathbf{x}^* + \mathbf{e}_k$, where $\mathbf{x}^*, \mathbf{x}^* + \mathbf{e}_k \in \mathcal{S}$. Taking the worst-case k, this establishes the stochastic comparison in inequality (3). □

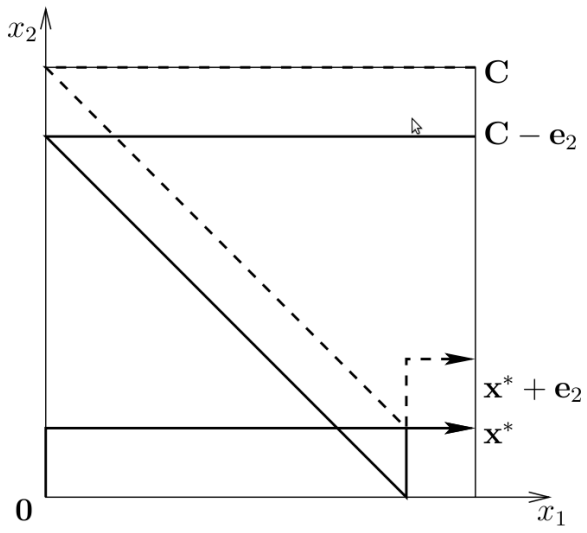

Figure 1: Illustration of (3) with two queues, $\mathbf{y} = \mathbf{C}$ and under the sequence of events $\{a_{10}^{(C_1)}, a_{21}^{(C_2-1)}, a_{02}, a_{01}^{(C_1)}\}$ ($a_{ij}^{(n)}$ reads a_{ij} for n times). The additional time $\tau(\mathbf{x}^*, \mathbf{x}^*+\mathbf{e}_2)$ provides a bound on $\tau(\mathbf{0}, \mathbf{C})$.

Proposition 2 provides a recursive framework to analyze $\tau(\mathbf{0}, \mathbf{C})$ in terms of the coupling times of trajectories having "distance one" at most, i.e., $\tau(\mathbf{x}^*, \mathbf{x}^*+\mathbf{e}_k)$. The fact that the trajectories from $\mathbf{x}^*$ and $\mathbf{x}^*+\mathbf{e}_k$ remain at distance one at each time step follows by Lemma 1, used in the proof of Proposition 2. As we show in the following, this decomposition of the problem is very useful because these trajectories appear easier to analyze than directly $\tau(\mathbf{0}, \mathbf{C})$.

When a single queue is considered, i.e., $M = 1$, the next proposition is known for the mean hitting time from state x to state zero, defined as $\mathbb{E}\min\{n : \phi^{(n)}(x, u_{1\to n}) = 0\}$ [11].

Proposition 3. *Assume $M = 1$ and $\rho \stackrel{\text{def}}{=} \lambda/\mu < 1$. Then,*

$$\mathbb{E}\min\{n : \phi^{(n)}(x, u_{1\to n}) = 0\} \leq \frac{1+\rho}{1-\rho}\, x. \tag{5}$$

For an $M/M/1/C$ queue, it is clear that the mean hitting time from C to 0 bounds from above its mean coupling time of the trajectories from states C and 0. This observation, together with previous proposition, will be exploited in Proposition 4.

Lemma 2. *Let $(\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n}))_{n\in\mathbb{N}}$ be the Markov chain under investigation when $\mathcal{S} \equiv \mathbb{N}^M$, i.e., $C_i = +\infty$, for all i. Let also $\tau_\infty(\mathbf{x}-\mathbf{e}_i, \mathbf{x})$ be as in (2) but when the state space of the Markov chain is $\mathcal{S} \equiv \mathbb{N}^M$. Then,*

$$\phi^{(n)}(\mathbf{x}, u_{1\to n})_i \leq_{st} \phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i,\ \forall i, n, \mathbf{x} \in \mathcal{S} \tag{6}$$

$$\tau(\mathbf{x}-\mathbf{e}_i, \mathbf{x}) \leq_{st} \tau_\infty(\mathbf{x}-\mathbf{e}_i, \mathbf{x}),\ \forall i, \mathbf{x}, \mathbf{x}-\mathbf{e}_i \in \mathcal{S}. \tag{7}$$

Proof. Relation (6) follows immediately by coupling the sample paths of both processes $(\phi^{(n)}(\mathbf{x}, u_{1\to n}))_{n\in\mathbb{N}}$ and $(\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n}))_{n\in\mathbb{N}}$ under the same sequence of events. Then, using (6) and the list of transitions in the proof of Lemma 1, one can easily verify that if $\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n}) = \phi_\infty^{(n)}(\mathbf{x}-\mathbf{e}_i, u_{1\to n})$, then $\phi^{(n)}(\mathbf{x}, u_{1\to n}) = \phi^{(n)}(\mathbf{x}-\mathbf{e}_i, u_{1\to n})$. □

Let Γ_i be an upper bound on the maximum mean rate in which jobs can arrive in queue i. For any network state, the minimal upper bound that one can choose is

$$\Gamma_i \stackrel{\text{def}}{=} \sum_{j=0}^{M} \gamma_{ji},\ \forall i, \tag{8}$$

which corresponds to the situation where all queues are non-empty. Let also

$$\rho_i \stackrel{\text{def}}{=} \Gamma_i/\mu_i,\ \forall i. \tag{9}$$

The following proposition provides a bound on the mean coupling time of the trajectories that start from states $\mathbf{x}$ and $\mathbf{x}-\mathbf{e}_k$.

Proposition 4. *Let a JQN be given such that $\rho_i < 1$, $\forall i$. Then, for any $\mathbf{x} \in \mathcal{S}$ and k_1 such that $\mathbf{x}-\mathbf{e}_{k_1} \in \mathcal{S}$,*

$$\mathbb{E}\tau(\mathbf{x}-\mathbf{e}_{k_1}, \mathbf{x}) \leq c \max_i \{x_i, b\}. \tag{10}$$

where

$$b \stackrel{\text{def}}{=} \max_i \tfrac{\rho_i}{1-\rho_i},\quad c \stackrel{\text{def}}{=} (1+\textstyle\sum_{i=1}^M V_{k_1}(i)) \max_i \alpha_i \tfrac{1+\rho_i}{1-\rho_i} \tag{11}$$

and $(V_{k_1}(1), \ldots, V_{k_1}(M))$ is the unique solution of the linear system

$$\begin{cases} V_{k_1}(i) = \sum_{j=1, j\neq i}^M \frac{\gamma_{ij}}{\sum_{k=0,k\neq j}^M \gamma_{jk}} V_{k_1}(j), & \forall i \neq k_1 \\ V_{k_1}(k_1) = 1 + \sum_{j=1, j\neq k_1}^M \frac{\gamma_{k_1 j}}{\sum_{k=0,k\neq j}^M \gamma_{jk}} V_{k_1}(j). \end{cases} \tag{12}$$

Proof. Since $\mathbb{E}\tau(\mathbf{x}-\mathbf{e}_{k_1}, \mathbf{x}) \leq \mathbb{E}\tau_\infty(\mathbf{x}-\mathbf{e}_{k_1}, \mathbf{x})$ by Lemma 2, we find a bound on $\mathbb{E}\tau_\infty(\mathbf{x}-\mathbf{e}_{k_1}, \mathbf{x})$ (recall that the subscript ∞ denotes that $C_i = +\infty$, for all i). Before time $\tau_\infty(\mathbf{x}-\mathbf{e}_{k_1}, \mathbf{x})$, Lemma 1 ensures that

$$\sum_i (\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi_\infty^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i) = 1. \tag{13}$$

Therefore, for $n > 0$, we define

$$I(n) \stackrel{\text{def}}{=} \begin{cases} \operatorname*{argmax}_{i=1,\ldots,M} \phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi_\infty^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i, & \text{if } \sum_i \phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi_\infty^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i = 1 \\ 0, & \text{if } \sum_i \phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi_\infty^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i = 0 \end{cases} \tag{14}$$

and $I(0) \stackrel{\text{def}}{=} k_1$. Note that $I(n)$ is well-defined because $\sum_i \phi^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i \in \{0,1\}$ for all n, and therefore $I(n)$ is the index i of the unique queue such that $\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi_\infty^{(n)}(\mathbf{x}-\mathbf{e}_k, u_{1\to n})_i = 1$ if such index exists, and zero otherwise. The process $(I(n))_{n\in\mathbb{N}}$ starts in k_1 and remains in k_1 until when one of cases ii).c and iii).b occur (these cases are defined in the proof of Lemma 1). For instance, case iii).b corresponds to a routing event $a_{k_1k_2}$, for some $k_2 > 0$ such that $\gamma_{k_1k_2} > 0$, when $\phi_\infty^{(n)}(\mathbf{x}, a_{1\to n})_{k_1} = 1$ and $\phi_\infty^{(n)}(\mathbf{x}, a_{1\to n})_{k_2} < C_{k_2} = \infty$. If case iii).b occurs at time n, then $I(n+1) = k_2$. If ii).c occurs at time n, then $I(n+1) = 0$, and it will remain to zero for all $n' > n+1$ because both trajectories from $\mathbf{x}$ and $\mathbf{x}-\mathbf{e}_{k_1}$ met. Therefore, if H is the random variable of the total number of jumps of the process $(I(n))_{n\in\mathbb{N}}$, then the ergodicity of the Markov chain $\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})$ ensures that $H \geq 1$ is finite almost surely.

By construction, it is clear that $\tau_\infty(\mathbf{x}-\mathbf{e}_{k_1}, \mathbf{x}) = \min\{n : I(n) = 0\}$. Therefore, in the following we study the mean hitting time to zero of the $\{0, 1, \ldots, M\}$-valued stochastic process $(I(n))_{n\in\mathbb{N}}$, i.e., $\mathbb{E}\min\{n : I(n) = 0\}$.

For $h \geq 1$, let $T(h)$ be the random variable "time where the process $(I(n))_{n\in\mathbb{N}}$ makes the h-th jump if it exists, and ∞ otherwise" and $T(0) \stackrel{\text{def}}{=} 0$.

Using the law of total expectation, we have

$$\begin{aligned}
\mathbb{E}\min\{n : I(n) = 0\} &= \mathbb{E}\left(\mathbb{E}(\min\{n : I(n) = 0\}|H)\right) && (15)\\
&= \sum_{h:\Pr(H=h)>0} \mathbb{E}(\min\{n : I(n) = 0\}|H = h)\Pr(H = h) && (16)\\
&= \sum_{h:\Pr(H=h)>0} \mathbb{E}(T(h)|H = h)\Pr(H = h) && (17)\\
&= \sum_{h:\Pr(H=h)>0} \mathbb{E}\left(\sum_{h'=1}^{h} T(h') - T(h'-1)|H = h\right)\Pr(H = h) && (18)\\
&= \sum_{h:\Pr(H=h)>0} \sum_{h'=1}^{h} \mathbb{E}\left(T(h') - T(h'-1)|H = h\right)\Pr(H = h). && (19)
\end{aligned}$$

We now derive a bound for $\mathbb{E}\left(T(h') - T(h'-1)|H = h\right)$.

Lemma 3. *For $t \leq h$,*

$$\mathbb{E}(T(t) - T(t-1)|H = h) \leq \max_i \alpha_i \frac{1+\rho_i}{1-\rho_i} \max\left\{x_i, \frac{\rho_i}{1-\rho_i}\right\}. \tag{20}$$

Proof. Between times $T(t-1)$ and $T(t)-1$, $I(n)$ is constant, which means that there exists i such that $\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i = \phi_\infty^{(n)}(\mathbf{x} - \mathbf{e}_{k_1}, u_{1\to n})_i + 1$ for all $n \in \{T(t-1), \ldots, T(t)-1\}$.

Assume $t = 1$. Since $T(0) = 0$ and using that $\rho_i < 1$ for all i, it is clear that $\mathbb{E}(T(1)|H = h)$ can be upper bounded, up to the multiplicative constant α_{k_i} (see (1)), by the mean hitting time from x_{k_1} to 0 of an $M/M/1$ queue with arrival rate Γ_{k_1} and service rate μ_{k_1}. This holds true by simply coupling such $M/M/1$ queue with the one-dimensional process $(\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_{k_1})_{n\in\mathbb{N}}$ in the instants where $(\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n}))_{n\in\mathbb{N}}$ moves along dimension k_1. The stochastic dominance of such $M/M/1$ queue holds because Γ_{k_1} is the maximum rate in which jobs can join queue k_1. The hitting time to zero corresponds to the occurrence of one of cases ii).c and iii).b, which trigger a jump of $I(n)$. Therefore, using also Proposition 3, we obtain

$$\mathbb{E}(T(1)|H = h) \leq \alpha_{k_1} \frac{1+\rho_{k_1}}{1-\rho_{k_1}} x_{k_1}. \tag{21}$$

Assume $t > 1$. Using the same coupling argument above, $\mathbb{E}(T(t)-T(t-1)|H = h)$ can be upper bounded by the mean hitting time from some initial state explicited in the following to 0 of an $M/M/1$ queue with arrival rate Γ_{k_t} and service rate μ_{k_t}, provided that $k_t = I(n)$ when $T(t-1) \leq n < T(t)$. The initial state of such $M/M/1$ queue must be greater than or equal to the random variable of the value of $\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_{k_t}$ at the time $n = T(t-1)$ of the $(t-1)$-th jump of $(I(n))_{n\in\mathbb{N}}$. We denote such random variable by $J(t-1)$. Conditioning on $J(t-1)$ and using again Proposition 3, we have

$$\begin{aligned}
\mathbb{E}(T(t) - T(t-1)|H = h) &\leq \sum_{x>0} \alpha_{k_t} \frac{1+\rho_{k_t}}{1-\rho_{k_t}} x \times \Pr(J(t-1) = x) && (22)\\
&= \alpha_{k_t} \frac{1+\rho_{k_t}}{1-\rho_{k_t}} \mathbb{E}J(t-1). && (23)
\end{aligned}$$

The following lemma provides a bound on $\mathbb{E}J(t-1)$ and concludes the proof of (20).

Lemma 4. $\mathbb{E}J(t-1) \leq \max\left\{x_{k_t}, \frac{\rho_{k_t}}{1-\rho_{k_t}}\right\}$.

Proof. For all t, we recall that we defined k_t as the value of $I(n)$ when $T(t-1) \leq n < T(t)$. First, we observe that $\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})$ is stochastically bounded by the vector $Y(n) \stackrel{\text{def}}{=} (Y_1(n), \ldots, Y_M(n))$ of the number of jobs in M independent $M/M/1$ queues at time n where each queue i has arrival rate Γ_i, service rate μ_i and it is started in x_i, i.e., $Y_i(0) = x_i$, for all i. This property is shown in [18] and follows again by coupling the

sample paths of both vector processes under the same sequence of events. Therefore, for any time n (random or not),

$$\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_i \leq_{st} Y_i(n),\ \forall i. \tag{24}$$

It is well-known, e.g., [1], that for the stable $M/M/1$ queue $Y_i(n)$,

$$\mathbb{E}Y_i(n) \leq \max\left\{Y_i(0), \frac{\rho_i}{1-\rho_i}\right\} \tag{25}$$

for any non-random time n. Using (24) in the former, we get

$$\mathbb{E}\,\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_{k_t} \leq \max\left\{x_{k_t}, \frac{\rho_{k_t}}{1-\rho_{k_t}}\right\} \tag{26}$$

for any non-random time n. The proof is thus concluded if we show that (26) holds true even when n is the random time $T(t-1)$, which is what we prove now. We recall that $J(t-1)$ is $\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_{k_t}$ when n is the time $T(t-1)$ of the $(t-1)$-th jump of $(I(n))_{n\in\mathbb{N}}$. Conditioning on $T(t-1)$, we have

$$\begin{aligned}
\mathbb{E}J(t-1) &= \mathbb{E}\phi_\infty^{(T(t-1))}(\mathbf{x}, u_{1\to T(t-1)})_{k_t} && (27)\\
&= \sum_n \mathbb{E}[\phi_\infty^{(n)}(\mathbf{x}, u_{1\to n})_{k_t} | T(t-1)=n]\Pr(T(t-1)=n) && (28)\\
&\overset{\text{using (24)}}{\leq} \sum_n \mathbb{E}[Y_{k_t}(n)|T(t-1)=n]\Pr(T(t-1)=n) && (29)\\
&\overset{\text{using (25)}}{\leq} \sum_n \max_i\left\{Y_i(0), \frac{\rho_i}{1-\rho_i}\right\}\Pr(T(t-1)=n) && (30)\\
&= \max_i\left\{Y_i(0), \frac{\rho_i}{1-\rho_i}\right\}. && (31)
\end{aligned}$$

In (30), we have used that $T(t-1)$ and $Y_{k_t}(n)$ are independent random variables: in fact, by construction, $T(t-1)$ is a random variable that only depends on some queue $k_{t-1} \neq k_t$ that hits zero (because either case ii).c or iii).b occurs at time $T(t-1)-1$) and all the M queues $Y_i(n)$'s are independent. □

□

Using (20) in (19), we obtain

$$\mathbb{E}\min\{n : I(n)=0\} \leq \max_i \alpha_i \frac{1+\rho_i}{1-\rho_i}\ \max\left\{x_i, \frac{\rho_i}{1-\rho_i}\right\}\sum_{h>0} h\Pr(H=h). \tag{32}$$

Therefore, it sufficies to understand the magnitude of $\mathbb{E}H = \sum_h h\Pr(H=h)$. Now, let $\tilde{I}(t)$ be the value of $(I(n))_{n\in\mathbb{N}}$ at the time where it makes the t-th jump. Since the jumps of $(I(n))_{n\in\mathbb{N}}$ are only due to occurrences of cases ii).c and iii).b, if $\tilde{I}(t)=j$ then the probability that $\tilde{I}(t+1)=i$ is $\frac{\gamma_{ji}}{\sum_{k=0,k\neq j}^M \gamma_{jk}}$, for all j,i. Given these jump probabilities, we can interpret the stochastic process $(\tilde{I}(t))_{t\in\mathbb{N}}$ as a discrete-time Markov chain where state zero is absorbing. Given this interpretation, we let $V_{k_1}(i)$ denote the mean number of visits that $(\tilde{I}(t))_{t\in\mathbb{N}}$ makes to state i when it is started in k_1. This means $\mathbb{E}H = \sum_{i=0}^M V_{k_1}(i)$. From the theory of absorbing Markov chains [14, Chapter 3], $\mathbb{E}H$ is called 'mean time to absorption' and it is known that the $V_{k_1}(i)$'s are given by the solution of (12). It is clear that $V_{k_1}(0)=1$ because $(\tilde{I}(t))_{t\in\mathbb{N}}$ remains in state zero after having visited (absorbing) state zero. Substituting this in (32) proves (10). □

With the above propositions, we are now in a position to state and prove our main result.

Theorem 2. *Let a JQN be given such that $\rho_i < 1$, $\forall i$. Then,*

$$\mathbb{E}\tau(\mathbf{0}, \mathbf{C}) \leq cM \sum_i \frac{\rho_i}{1-\rho_i} \times \sum_i C_i \tag{33}$$

where c is given by (11).

Proof. Recall that $u_{-n+1\to 0}$ is a sequence of n independent events from time $-n+1$ to time 0.

For simplicity of notation, let $\mathbf{x}^*(\mathbf{y}) \stackrel{\text{def}}{=} \phi^{(n^*)}(\mathbf{y}, u_{-n^*+1\to 0}) = \phi^{(n^*)}(\mathbf{0}, u_{-n^*+1\to 0})$ where $n^* \stackrel{\text{def}}{=} n^*(\mathbf{y})$ is the backward coupling time between $\mathbf{0}$ and $\mathbf{y}$, i.e., $\min\{n \in \mathbb{N} \,|\, \phi^{(n)}(\mathbf{0}, u_{-n+1\to 0}) = \phi^{(n)}(\mathbf{y}, u_{-n+1\to 0})\}$. By definition, $n^*(\mathbf{y})$ has the same distribution as $\tau(\mathbf{y}, \mathbf{0})$, and therefore $n^* < \infty$ almost surely by Theorem 1. We have

$$\begin{aligned}
\Pr(x_i^*(\mathbf{y}) \geq x_i) &\leq \Pr(x_i^*(\mathbf{C}) \geq x_i) && (34)\\
&= \lim_{n\to\infty} \Pr(\phi^{(n)}(\mathbf{0}, u_{1\to n})_i \geq x_i) && (35)\\
&\leq \lim_{n\to\infty} \Pr(\phi_\infty^{(n)}(\mathbf{0}, u_{1\to n})_i \geq x_i). && (36)
\end{aligned}$$

The first inequality comes from the fact that $\phi(\cdot)$ is monotone; the following equality comes from Theorem 1; the last inequality follows from (6).

Using Propositions 2 and 4, if $\rho_i < 1$ for all i, then

$$\begin{aligned}
\mathbb{E}\tau(\mathbf{0}, \mathbf{y}+\mathbf{e}_i) - \mathbb{E}\tau(\mathbf{0}, \mathbf{y}) &\leq \mathbb{E}[\mathbb{E}[\max_k \tau(\mathbf{x}^*(\mathbf{y}), \mathbf{x}^*(\mathbf{y}) + \mathbf{e}_k) | \mathbf{x}^*(\mathbf{y})]] && (37)\\
&\leq \sum_k \mathbb{E}[\mathbb{E}[\tau(\mathbf{x}^*(\mathbf{y}), \mathbf{x}^*(\mathbf{y}) + \mathbf{e}_k) | \mathbf{x}^*(\mathbf{y})]] && (38)\\
&\leq \sum_k \sum_{\mathbf{x}\in\mathbb{N}^M} \mathbb{E}[\tau(\mathbf{x}^*(\mathbf{y}), \mathbf{x}^*(\mathbf{y}) + \mathbf{e}_k) | \mathbf{x}^*(\mathbf{y}) = \mathbf{x}] \Pr(\mathbf{x}^*(\mathbf{y}) = \mathbf{x}) && (39)\\
&\stackrel{\text{using (10)}}{\leq} M \sum_{\mathbf{x}\in\mathbb{N}^M} c \max_i\{x_i, b\} \Pr(\mathbf{x}^*(\mathbf{y}) = \mathbf{x}) && (40)\\
&= Mc \max\{b, \mathbb{E}[\max_i x_i^*(\mathbf{y})]\} && (41)\\
&= Mc \max\left\{b, \sum_{x\geq 1} \Pr(\max_i x_i^*(\mathbf{y}) \geq x)\right\} && (42)\\
&\stackrel{\text{using (36)}}{\leq} Mc \max\left\{b, \sum_{x\geq 1} \lim_{n\to\infty} \Pr(\max_i \phi_\infty^{(n)}(\mathbf{0}, u_{1\to n})_i \geq x)\right\} && (43)\\
&\leq Mc \max\left\{b, \sum_{x\geq 1} \lim_{n\to\infty} \Pr\Big(\sum_i \phi_\infty^{(n)}(\mathbf{0}, u_{1\to n})_i \geq x\Big)\right\} && (44)\\
&\leq Mc \max\left\{b, \sum_i \frac{\rho_i}{1-\rho_i}\right\} && (45)\\
&= Mc \sum_i \frac{\rho_i}{1-\rho_i}. && (46)
\end{aligned}$$

In (45), we have used that the mean stationary number of jobs in queue i of a Jackson network with infinite buffers is geometric. Now, for $\mathbf{y} = \mathbf{C} - \mathbf{e}_i$ and following the recursion along dimension i we have

$$\mathbb{E}\tau(\mathbf{0}, \mathbf{C}) \leq \mathbb{E}\tau(\mathbf{0}, \mathbf{C} - C_i\mathbf{e}_i) + Mc \textstyle\sum_i \frac{\rho_i}{1-\rho_i} C_i. \quad (47)$$

Solving the recursion along each dimension, we get (33). □

Theorem 2 provides sufficient conditions to characterize the qualitative behavior of the mean coupling time of JQNs when the network size, i.e., the number of queues and the buffer sizes, varies. In the following remark, we derive such qualitative behavior.

Remark 1. *We recall that the $V_{k_1}(i)$'s, as defined in (12), are interpreted as the mean number of visits performed to each queue of our JQN by a job that starts visiting the network from queue k_1; see, e.g., [7, Chapter 7]. Suppose we are given a sequence of JQNs indexed by r where $M^{(r)}$ and $\mathbf{C}^{(r)}$ are increasing in*

$r \in \mathbb{N}$. *Furthermore, in each JQN of the sequence, assume that* $\frac{\max_i \alpha_i^{(r)}}{M^{(r)}}$, $\max_i \rho_i^{(r)}$, $V_i^{(r)}$ *(solution of* (12)*) for all* $i, j \geq 0$, *are uniformly bounded by a constant independent of the sequences* $(M^{(r)})_{r\geq 0}$ *and* $(\mathbf{C}^{(r)})_{r\geq 0}$. *As a consequence,* $c = O((M^{(r)})^2)$, *and Theorem 2 shows that the mean coupling time of the* r*-th JQN is*

$$O\Big((M^{(r)})^4 \sum_i C_i^{(r)}\Big). \tag{48}$$

Formula (48) *shows that the coupling time of JQNs is polynomial in both* M *and* $\mathbf{C}$. *This improves the bounds developed in [3] (a comparison with those bounds follows below).*

The following remark comments on the assumption of Theorem 2.

Remark 2. *For a network with arbitrary topology and service rates, the assumption* $\rho_i = \Gamma_i/\mu_i < 1$ *corresponds to what can be called an "hyper-stable" queue. Indeed, in such a case, the arrival rate in queue* i *is always smaller than its service rate, even when all input queues in* i *are not empty and send jobs to* i *at full rate. This does not mean that the network is in a light-load regime. In fact, by varying the external arrival rates and the routing probabilities of some downstream queues, each queue can have any load between 0 and 1 within this assumption. From a technical standpoint, this assumption implies that the* drift *along each dimension* i *of the state space of the underlying Markov chain transitions can be bounded uniformly (from above) by a positive constant, i.e.,* Γ_i. *This structure is essentially exploited in the proof of Proposition 4. This bounded drift assumption suggests that our main result can be proved using some Lyapunov function on the expected time to empty the system. However, this approach cannot work as we discuss in Section 3.1.*

3.1 Time to empty the system

It is easy to see that $\tau(\mathbf{0}, \mathbf{C})$ can be stochastically bounded from above by the hitting time from $\mathbf{C}$ to $\mathbf{0}$, i.e., $\min\{n : \phi^{(n)}(\mathbf{C}, u_{1\to n}) = \mathbf{0}\}$. In turn, using a simple coupling argument, we have

$$\min\{n > 0 : \phi^{(n)}(\mathbf{0}, u_{1\to n}) = \mathbf{0}\} \leq_{st} \min\{n : \phi^{(n)}(\mathbf{C}, u_{1\to n}) = \mathbf{0}\} \tag{49}$$

and, using Kac's lemma, $\mathbb{E}\min\{n > 0 : \phi^{(n)}(\mathbf{0}, u_{1\to n}) = \mathbf{0}\} = 1/\pi(\mathbf{0})$, where $\pi(\mathbf{0})$ denotes the stationary probability of being in state $\mathbf{0}$. However, it is known that $1/\pi(\mathbf{0})$ grows exponentially in M: in fact, if the C_i's are sufficiently large and the overall mean arrival rate to i, say λ_i, is such that $\lambda_i < \mu_i$, then $\pi(\mathbf{0}) \approx \prod_{i=1}^{M}(1 - \lambda_i/\mu_i)$ [8].

Remark 3. *The argument above shows that we cannot use an argument based on a Lyapunov function and on the expected time to empty the system to prove polynomial bounds on* $\mathbb{E}\tau(\mathbf{0}, \mathbf{C})$ *when* M *and the* C_i*'s increase, even when the drift to* $\mathbf{0}$ *of the underlying Markov chain is uniformly bounded as in the assumption of Theorem 2.*

3.2 Comparison with [3]

The main difference of our bound (33) with respect to the ones in [3] is that (33) is of the order of a polynomial in both M and $\mathbf{C}$ (see (48)). In other words, it is not exponential in either M or the C_i's.

For the case of JQN with cycles, Proposition 4.3 in [3] gives some conditions to prove that the coupling time of JQN is linear in the buffer size of *a single* queue. In particular, the constant of proportionality is not specified and, most importantly, the number of queues M is considered as a constant (in contrast with our approach). If *all* the buffer sizes are allowed to grow, these conditions are stronger than the ones in Theorem 2. In fact, there it is required hyperstability (as we do) plus the condition on the network topology that each queue can either receive jobs from outside or send jobs to outside, i.e., $\gamma_{i0} > 0$ or $\gamma_{0i} > 0$, for all i.

4 Bound on mixing time

The mixing time of a discrete-time Markov chain $(X_n)_{n\in\mathbb{N}}$ with state space S is the time n needed for the measure of X_n to be close to the stationary measure up to a precision threshold. This notion is very useful to

get probabilistic guarantees for Monte Carlo simulation. More precisely, let P be the transition matrix of the chain and let $d(n) \stackrel{\text{def}}{=} \max_{x \in S} ||P^n(x, \cdot) - \pi(\cdot)||_{TV}$, where $|| \cdot ||_{TV}$ is the total variation distance. The *mixing time* with uncertainty ϵ, $m(\epsilon)$, is the time it takes for the total variation distance between the distribution of the chain at time n and the stationary distribution π, to be less than or equal to ϵ, uniformly on the initial state. Namely,

$$m(\epsilon) \stackrel{\text{def}}{=} \min\{n : d(n) \leq \epsilon\}.$$

The mixing time is usually defined with uncertainty 1/4: $T_{mix} \stackrel{\text{def}}{=} m(1/4)$ (see [17]).

If we consider the Markov chain defined by a monotone QN with finite buffer, then its mixing time is bounded from above by its coupling time. This is stated in the following proposition, whose proof is based on rather classical arguments that relate the mixing and the coupling time.

Proposition 5. *Let a monotone QN with buffer vector* $\mathbf{C}$ *be given. Then, for all* $\epsilon > 0$

$$m(\epsilon) \leq 4\lceil \log_2 \frac{1}{\epsilon} \rceil \mathbb{E}\tau(\mathbf{0}, \mathbf{C}). \tag{50}$$

Proof. Theorem 5.2 in [17] says that $||P^n(\mathbf{x}, \cdot) - P^n(\mathbf{y}, \cdot)||_{TV} \leq \Pr(\tau(\mathbf{x}, \mathbf{y}) > n)$. From this point, let $s(n)$ be the total variation between to transient measures of the chain. We have

$$\begin{array}{rcl} s(n) = & \max_{\mathbf{x},\mathbf{y} \in \mathcal{S}} ||P^n(\mathbf{x}, \cdot) - P^n(\mathbf{y}, \cdot)||_{TV} \\ \leq & \max_{\mathbf{x},\mathbf{y} \in \mathcal{S}} \Pr(\tau(\mathbf{x}, \mathbf{y}) > n) \\ = & \Pr(\tau(\mathbf{0}, \mathbf{C}) > n), \end{array} \tag{51}$$

where the last equality comes from the monotonicity of the chain. Now, the mixing can also be bounded since by definition $d(n) \leq s(n)$,

$$\begin{array}{rcl} m(\epsilon) = & \min\{n : d(n) \leq \epsilon\} \\ \leq & \min\{n : \Pr(\tau(\mathbf{0}, \mathbf{C}) > n) \leq \epsilon\} \\ \leq & \min\{n : \mathbb{E}\tau(\mathbf{0}, \mathbf{C})/n \leq \epsilon\} \\ = & \mathbb{E}\tau(\mathbf{0}, \mathbf{C})/\epsilon, \end{array} \tag{52}$$

where the penultimate inequality is the Markov inequality. By replacing ϵ by 1/4, one gets $T_{mix} \leq 4\mathbb{E}\tau(\mathbf{0}, \mathbf{C})$. Again, for an arbitrary integer k, $d(km(\epsilon)) \leq s(km(\epsilon))$. Since s is sub-multiplicative (see [17]), $s(km(\epsilon)) \leq s(m(\epsilon))^k$. By definition of s, $s(n) \leq 2d(n)$. Therefore, $s(m(\epsilon))^k \leq (2d(m(\epsilon))^k \leq (2\epsilon)^k$. In total, we get $d(km(\epsilon)) \leq (2\epsilon)^k$. By taking $\epsilon = 1/4$ the later inequality provides $d(kT_{mix}) \leq 2^{-k}$ and $m(\epsilon) \leq \lceil \log_2 \epsilon^{-1} \rceil T_{mix}$. □

By means of Proposition 5, thus, it is possible to exploit the results on the coupling time presented in Section 3 to bound the mixing time.

5 Queueing networks with blocking

The proof technique presented in Section 3 can be applied to other types of QNs with finite buffers, provided that they remain monotone and hyper-stable. For example, the same approach can be used for other types of QNs

- with other types of mechanisms to deal with full buffers (e.g,. with blocking),
- with state-dependent routing (e.g., join-the-shortest-queue),
- with stations having service rates depending on the number of jobs in their queues (e.g., $-/M/k/C$ queues), and
- cases where jobs belong to multiple classes.

In all these cases, the coupling time can be shown to be polynomial both in the number of queues and buffer sizes. Bounds on the mean mixing time trivially follow by using Proposition 5, which only requires the monotonicity of the chain.

Here, we provide a detailed treatment of the blocking case. There exist many blocking mechanisms, and the most classical ones include blocking after service, blocking before service, repetitive service or recirculate blocking. They are all similar in terms of coupling and mixing times. We consider the case with repetitive service, where upon completion of service at queue i, a job is sent to queue j with probability p_{ij} and if j has no available slots, queue i repeats the service for that job. Afterwards, a new target queue is selected according to the routing probabilities of i in a i.i.d. manner. As mentioned before, this type of blocking is called Repetitive-Service (RS) blocking, and even for this class of JQNs no product-form is known for the stationary distribution; see [5] for a background and further details.

Remark 4. *The evolution of a JQN with RS blocking is thus identical to the model introduced in Section 2 except for routing events to queues that are full, for which the state remains unchanged (a state is still represented by the number of jobs in each queue).*

Furthermore, quantity Γ_i, defined in (8), still represents the minimal upper bound on the mean arrival rate of jobs in queue i with respect to any network state. We also observe that JQNs with RS blocking are monotone (as shown in [3]).

Proposition 6. *Let a JQN with RS blocking be given. For all i and $\mathbf{y} \in \mathcal{S} : y_i > 0$, (3) holds true.*

Proof. For any possible event $a \in \mathcal{A}$, $k = 1, \ldots, M$, and any state $\mathbf{y}$ such that $\mathbf{y}, \mathbf{y} - \mathbf{e}_k \in \mathcal{S}$, (4) still holds true. In fact, for every $\mathbf{y} \in \mathcal{S}$ and k such that $\mathbf{y} - \mathbf{e}_k \in \mathcal{S}$, cases i.a)–ii.c) of the proof of Proposition 2 holds for JQNs with RS blocking because their dynamics are equivalent to the ones of Section 2 (within these cases). For each routing events $a = a_{ij}$, we have the following cases

iii.a) If $\mathbf{y} : 1 < y_i \leq C_i$ and $k = i$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

iii.b) If $\mathbf{y} : y_i = 1$ and $k = i$ and $y_j < C_j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_j$

iii.c') If $\mathbf{y} : y_i = 1$ and $k = i$ and $y_j = C_j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

iii.d) If $\mathbf{y} : y_j < C_j$ and $k = j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

iii.e') If $\mathbf{y} : y_j = C_j$ and $k = j$ and $y_i \neq 0$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_j$

iii.f) If $\mathbf{y} : y_j = C_j$ and $k = j$ and $y_i = 0$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$

iii.g) If $k \neq i$ and $k \neq j$, then $\phi(\mathbf{y}, a) = \phi(\mathbf{y} - \mathbf{e}_k, a) + \mathbf{e}_k$,

i.e., only cases iii.c) and iii.e) change. Therefore, (4) holds true and, at time $\tau(\mathbf{0}, \mathbf{y} - \mathbf{e}_i)$, both trajectories from $\mathbf{0}$ and $\mathbf{y}$ will be at distance one. □

The proofs of Proposition 4 and Theorem 2 are based on a stochastic dominance property of JQNs with infinite buffers with respect to their finite counterparts (see Lemma 2). In JQNs with RS blocking, this property is lost and thus those proofs cannot be applied directly to obtain the same bound. However, a quadratic bound in the C_i's can be derived as we show in the following.

The proof of the following proposition proceeds on the same lines of the proof of Proposition 4.

Proposition 7. *Let a JQN with RS blocking be given such that $\rho_i \stackrel{\text{def}}{=} \Gamma_i/\mu_i < 1$, $\forall i$. Then, for any $\mathbf{x} \in \mathcal{S}$ and k_1 such that $\mathbf{x} - \mathbf{e}_{k_1} \in \mathcal{S}$,*

$$\mathbb{E}\tau(\mathbf{x} - \mathbf{e}_{k_1}, \mathbf{x}) \leq c \max_i C_i. \tag{53}$$

where c is

$$c \stackrel{\text{def}}{=} (1 + \textstyle\sum_{i=1}^{M} V_{k_1}(i)) \max_i \alpha_i \frac{1+\rho_i}{1-\rho_i} \tag{54}$$

and $(V_{k_1}(1), \ldots, V_{k_1}(M))$ is the unique solution of the linear system

$$\begin{cases} V_{k_1}(i) = \sum_{j=1, j\neq i}^{M} \max\left\{ \frac{\gamma_{ji}}{\sum_{k=0,k\neq j}^{M} \gamma_{jk}}, \frac{\gamma_{ji}}{\sum_{k=0,k\neq j}^{M} \gamma_{kj}} \right\} V_{k_1}(j), & \forall i \neq k_1 \\ V_{k_1}(k_1) = 1 + \sum_{j=1, j\neq k_1}^{M} \max\left\{ \frac{\gamma_{jk_1}}{\sum_{k=0,k\neq j}^{M} \gamma_{jk}}, \frac{\gamma_{jk_1}}{\sum_{k=0,k\neq j}^{M} \gamma_{kj}} \right\} V_{k_1}(j). \end{cases} \tag{55}$$

Proof. Similarly to (14), let

$$I(n) \stackrel{\text{def}}{=} \begin{cases} \underset{i=1,\dots,M}{\operatorname{argmax}} \phi^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i, & \text{if } \sum_i \phi^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i = 1 \\ 0, & \text{if } \sum_i \phi^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i = 0 \end{cases} \tag{56}$$

for $n > 0$ and $I(0) \stackrel{\text{def}}{=} k_1$.[1] Note that (56) is well-defined because $\sum_i \phi^{(n)}(\mathbf{x}, u_{1\to n})_i - \phi^{(n)}(\mathbf{x}-\mathbf{e}_{k_1}, u_{1\to n})_i \in \{0,1\}$. By construction, it is clear that $\tau(\mathbf{x}-\mathbf{e}_{k_1}, \mathbf{x}) = \min\{n : I(n) = 0\}$. Therefore, in the following we study $\mathbb{E}\min\{n : I(n) = 0\}$.

For $h \geq 1$, let $T(h)$ be as in the proof of Proposition 7 be the random variable "time where the process $(I(n))_{n\in\mathbb{N}}$ makes the h-th jump if it exists, and ∞ otherwise", and let $T(0) = 0$.

Using the same argument in (19), we have

$$\mathbb{E}\min\{n : I(n) = 0\} \quad = \quad \sum_{h:\Pr(H=h)>0} \sum_{h'=1}^{h} \mathbb{E}\left(T(h') - T(h'-1) | H = h\right) \Pr(H = h), \tag{57}$$

and therefore we derive a bound for $\mathbb{E}\left(T(h') - T(h'-1)|H = h\right)$ and on $\Pr(H = h)$, where we recall that H is the random variable of the total number of jumps of the process $(I(n))_{n\in\mathbb{N}}$.

As in Lemma 3, $\mathbb{E}(T(t) - T(t-1)|H = h)$ can be upper bounded by the mean hitting time from some state to 0 of an $M/M/1/C_{k_t}$ with arrival rate Γ_{k_t} and service rate μ_{k_t}, provided that $k_t = I(n)$ when $T(t-1) \leq n < T(t)$. Using Proposition 3, in the worst case we trivially have

$$\mathbb{E}(T(t) - T(t-1)|H = h) \leq \max_i \alpha_i C_i \frac{1+\rho_i}{1-\rho_i}, \tag{58}$$

and substituting in (57) we obtain

$$\mathbb{E}\min\{n : I(n) = 0\} \quad = \quad \max_i \alpha_i C_i \frac{1+\rho_i}{1-\rho_i} \mathbb{E}(H), \tag{59}$$

Now, let $\tilde{I}(t)$ be the value of $(I(n))_{n\in\mathbb{N}}$ at the time of its t-th jump.

The jumps of $(I(n))_{n\in\mathbb{N}}$ to state $k > 0$ are only due to occurrences of cases iii).b and iii).c' (defined in the proof of Proposition 6), which means that at the time $n+1$ of a jump of I we have either $\phi^{(n)}(\mathbf{x}, u_{1\to n})_{I(n)} = 1$ or $\phi^{(n)}(\mathbf{x}, u_{1\to n})_{I(n)} = C_{I(n)}$. Therefore, if $\tilde{I}(t) = j$ and $n_t + 1$ is the time of the t-th jump then

$$\Pr(\tilde{I}(t+1) = i | \tilde{I}(t) = j \text{ and } \phi^{(n_t)}(\mathbf{x}, u_{1\to n_t})_{I(n_t)} = 1) \quad = \quad \frac{\gamma_{ji}}{\sum_{k=0,k\neq j}^{M} \gamma_{jk}} \tag{60}$$

$$\Pr(\tilde{I}(t+1) = i | \tilde{I}(t) = j \text{ and } \phi^{(n_t)}(\mathbf{x}, u_{1\to n_t})_{I(n_t)} = C_{I(n_t)}) \quad = \quad \frac{\gamma_{ij}}{\sum_{k=0,k\neq j}^{M} \gamma_{kj}} \tag{61}$$

for all i, j. Let $V_{k_1}(i)$ denote the mean number of visits that $(\tilde{I}(t))_{t\in\mathbb{N}}$ makes to state i when it is started in k_1. The above conditional jump probabilities imply

$$\Pr(\tilde{I}(t+1) = i | \tilde{I}(t) = j) \quad \leq \quad \max\left\{ \frac{\gamma_{ji}}{\sum_{k=0,k\neq j}^{M} \gamma_{jk}}, \frac{\gamma_{ij}}{\sum_{k=0,k\neq j}^{M} \gamma_{kj}} \right\}, \tag{62}$$

and we can upper bound $\mathbb{E}H$ easily as done in the proof of Proposition 4 by summing the $V_{k_1}(i)$'s. It is clear that $V_{k_1}(0) = 1$, because $(I(n))_{n\in\mathbb{N}}$ remains in state zero after having visited zero, and upper bounds on the $V_{k_1}(i)$'s are given by (55), which uses (62). Substituting this in (57) proves (53). □

We are now in a position to prove the following result.

[1] The main difference with (14) is that (56) is defined in terms of $\phi(\cdot)$ rather than $\phi_\infty(\cdot)$. This is necessary because in the case of RS blocking one can see that $\phi(\mathbf{x}, a) \not\leq_{st} \phi_\infty(\mathbf{x}, a)$, and the argument in the proof of Proposition 4 does not extend immediately here.

Theorem 3. *Let a JQN with RS blocking be given such that $\rho_i < 1$, $\forall i$. Then, $\mathbb{E}\tau(\mathbf{0}, \mathbf{C}) \leq Mc \max_i C_i \sum_i C_i$, where c is given by* (54).

Proof. Using Propositions 6 and 7, we have

$$\begin{aligned} \mathbb{E}\tau(\mathbf{0}, \mathbf{y} + \mathbf{e}_i) - \mathbb{E}\tau(\mathbf{0}, \mathbf{y}) &\leq \mathbb{E}[\mathbb{E}[\max_k \tau(\mathbf{x}^*(\mathbf{y}), \mathbf{x}^*(\mathbf{y}) + \mathbf{e}_k)|\mathbf{x}^*(\mathbf{y})]] & (63) \\ &\leq \sum_k \mathbb{E}[\mathbb{E}[\tau(\mathbf{x}^*(\mathbf{y}), \mathbf{x}^*(\mathbf{y}) + \mathbf{e}_k)|\mathbf{x}^*(\mathbf{y})]] & (64) \\ &\leq Mc \max_j C_j. & (65) \end{aligned}$$

Following the recursion along dimension i, we have

$$\mathbb{E}\tau(\mathbf{0}, \mathbf{C}) \leq \mathbb{E}\tau(\mathbf{0}, \mathbf{C} - C_i\mathbf{e}_i) + Mc\, C_i \max_j C_j. \quad (66)$$

Solving the recursion along each dimension, we finally find $\mathbb{E}\tau(\mathbf{0}, \mathbf{C}) \leq Mc \max_j C_j \sum_i C_i$. □

In the same conditions of Remark 1 and the ones of Theorem 3, we conclude that the qualitative behavior of the mean coupling time when both M and $\mathbf{C}$ increase is

$$O\Big((M^{(n)})^3 \max_i C_i^{(n)} \sum_i C_i^{(n)}\Big). \quad (67)$$

6 Conclusions

We have proposed an approach to obtain bounds on the coupling and mixing times of a class of Markovian queueing networks, which are related to the efficiency of sampling from their stationary distribution in an exact or approximate manner, respectively. Our results give conditions to show that the coupling time of Jackson queueing networks with finite buffers and arbitrary topology grows slowly (polynomially) when both the number of queues and the size of all buffers increase. Our bounds significantly improve the best bounds known in the literature and extend the ones found for acyclic Jackson QNs in [3, 21]. We have also shown that the mixing time of finite Jackson QNs behaves in a similar way up to a factor depending on the accuracy of the desired samples. Under minor variations in the proof-technique used, we extended our approach to a class of queueing networks with blocking obtaining similar (polynomial in M and the C_i's) bounds. Other possible extensions include monotone networks where i) routing can be state-dependent, e.g., join-the-shortest-queue, ii) stations have service rates depending on the number of jobs in their queues, e.g., $-/M/k/C$ queues, and iii) cases where jobs belong to multiple classes. Again, these follow by adapting (with minor variations) the proofs of Propositions 2 and 4 and Theorem 2 or 3. In conclusion, our bounds promote both CFTP and Monte Carlo methods as efficient tools for evaluating the stationary performance of monotone queueing networks. While the former produces exact samples, the latter is more efficient, and one can choose among these two techniques based upon which properties best suit one's needs.

References

[1] J. Abate and W. Whitt. Transient Behavior of the M/M/1 Queue: Starting at the Origin *Queueing Syst.*, 2(1):41–65, 1987.

[2] C. Alexopoulos and D. Goldsman. To batch or not to batch? *ACM Trans. Model. Comput. Simul.*, 14(1):76–114, 2004.

[3] S. Andradóttir and M. Hosseini-Nasab. Efficiency of time segmentation parallel simulation of finite markovian queueing networks. *Operation Research*, 51(2):272–280, 2003.

[4] J. Anselmi and B. Gaujal. On the efficiency of perfect simulation in monotone queueing networks. *SIGMETRICS Perform. Eval. Rev, ACM*, 39(2):56–58, 2011.

[5] S. Balsamo, V. de Nitto Personé, and R. Onvural. *Analysis of queueing networks with blocking*. International series in operations research and management science. Kluwer, 2001.

[6] F. Baskett and K.M. Chandy and R. Muntz and F.G. Palacios Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22(2):248–260, 1975.

[7] U. Narayan Bhat. *An Introduction to Queueing Theory: Modeling and Analysis in Applications*. Birkhauser Verlag, 2008.

[8] G. Bolch, S. Greiner, H. de Meer, and K. Trivedi. *Queueing Networks and Markov Chains*. Wiley-Interscience, 2005.

[9] P. Bremaud. *Markov Chains, Gibbs Fields, Monte Carlo Simulation and Queues*. Texts in Applied Mathematics. Springer-Verlag, Berlin-Heidelberg, 1999.

[10] W.-L. Chen and C. A. O'Cinneide. Towards a polynomial-time randomized algorithm for closed product-form networks. *ACM Trans. Model. Comput. Simul.*, 8(3):227–253, 1998.

[11] J. Dopper, B. Gaujal, and J.-M. Vincent. Bounds for the coupling time in queueing networks perfect simulation. In *Celebration of the 100th anniversary of Markov*, pages 117–136, 2006.

[12] J. R. Jackson. Job shop-like queueing systems. *Management Sci.*, 10,131, 1963.

[13] F. Kelly. *Reversibility and Stochastic Networks*. 1979.

[14] J.G. Kemeny and J.L. Snell. *Finite Markov chains*. VanNostrand, University series in undergraduate mathematics, 1969.

[15] S. Kijima and T. Matsui. Approximation algorithm and perfect sampler for closed jackson networks with single servers. *SIAM J. Comput.*, 38(4):1484–1503, 2008.

[16] S. Kijima and T. Matsui. Randomized approximation scheme and perfect sampler for closed jackson networks with multiple servers. *Annals OR*, 162(1):35–55, 2008.

[17] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2008.

[18] W. A. Massey. An Operator Analytic Approach to the Jackson Network. *Journal of Applied Probability*, 21:379–393, 1984.

[19] A. Müller and D. Stoyan. *Comparison methods for stochastic models and risks*. Wiley, 2002.

[20] J. G. Propp and D. B. Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Rand. Struct. Alg.*, 9(1-2):223–252, 1996.

[21] J.-M. Vincent. Perfect generation, monotonicity and finite queueing networks. In *IEEE QEST*, page 319, 2008.

[22] W. Whitt. The efficiency of one long run versus independent replication in steady-state simulation. *Management Sci.*, 37(6):645–666, 1991.