Simulation Modelling Practice and Theory 37 (2013) 70–78

Contents lists available at SciVerse ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Rare event simulation of non-Markovian queueing networks using RESTART method

José Villén-Altamirano [a,*], Manuel Villén-Altamirano [b]

[a] Department of Matemática Aplicada, Technical University of Madrid, Calle Arboleda s/n, 28031 Madrid, Spain
[b] Department of Tecnología Electrónica, University of Malaga, Campus de Teatinos s/n, 29071 Malaga, Spain

ARTICLE INFO

Article history:
Received 14 February 2013
Received in revised form 29 May 2013
Accepted 31 May 2013
Available online 28 June 2013

Keywords:
Variance reduction
Queueing systems
Rare event
Splitting simulation
RESTART

ABSTRACT

RESTART is an accelerated simulation technique that allows the probabilities of rare events to be evaluated. In this method, a number of simulation retrials are performed when the process enters regions of the state space where the chance of occurrence of a rare event of interest is higher. These regions are defined by means of a function of the system state called the importance function. An appropriate choice of the importance function is crucial for the effective application of RESTART because, although the rare event estimator is unbiased for any importance function, the acceleration achieved is closely dependent on the selected function. Formulas for obtaining suitable importance functions to estimate overflow probabilities, previously provided for Jackson networks, are extended here to non-Markovian queueing networks. This extension is made by introducing an innovative concept, the effective load of a node, defined as the actual load of a node of a Jackson network which has a similar queue length distribution. The formulas are tested in four network topologies, ranging from a two-node network with strong feedback to a 15-node network with multiple feedbacks, with different interarrival and service time distributions. The paper shows how probabilities of rare events are accurately estimated in all the tested cases with short computational time. The large variety of cases simulated suggests that the proposed importance function may be suitable for many other queueing networks.

1. Introduction

Monte Carlo simulation plays an increasingly important role in the design and analysis of communications systems and networks, as they become more complex and the performance requirements more demanding. In many settings, evaluation of these requirements involves the estimation of rare event probabilities. For these estimations the development of efficient acceleration methods are necessary because crude simulations (i.e., simulations without any acceleration method) require prohibitively long execution times. This development is a challenging task, as was pointed out in [6]: "Efficient simulation of queueing networks is amongst the more difficult open problems in simulation".

One acceleration method is importance sampling (see [9] for a review). The basic idea behind this approach is to alter the probability measure governing events so that the rare event of interest occurs more often. One drawback of this technique is the difficulty of simulating large systems or systems with feedback. Examples of the difficulties that must be dealt with even in simple non-Markovian networks can be seen in [5].

Another method is RESTART (REpetitive Simulation Trials After Reaching Thresholds), introduced in [12], which has a precedent of much more limited scope in the splitting technique described in [7]. In RESTART a more frequent occurrence

* Corresponding author. Tel.: +34 913367525.
E-mail addresses: jvillen@eui.upm.es (J. Villén-Altamirano), manolo.villen@uma.es (M. Villén-Altamirano).

1569-190X/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.simpat.2013.05.012

of the rare event of interest is achieved by performing a number of simulation retrials when the process enters regions of the state space where the importance is greater, i.e., regions where the chance of occurrence of the rare event is higher. These regions, called importance regions, are defined by comparing the value taken by a function of the system state, the importance function, with certain thresholds. A theoretical analysis yielding the variance of the estimator, the gain obtained and the optimal values for thresholds and the number of retrials was made in [13]. The key point for achieving a high acceleration gain when RESTART is applied is the choice of the importance function. This function can be obtained heuristically or by means of analytical approximations, given that the estimator is unbiased for any importance function, as proven in [13] and corroborated in the multiple applications of RESTART made by many authors (see [14] and references therein). The only criterion for assessing the goodness of the choice is the acceleration achieved. In [8] it was pointed out that "in the case of multidimensional state spaces, good choices of the importance function for splitting are crucial, and are definitely non-trivial to obtain in general" and in [4] that "Although RESTART has been shown to be an efficient and flexible simulation technique for simple networks, its applicability for complex ones is still a challenging problem."

This paper deals with the problem of obtaining formulas for the importance function in non-Markovian networks to estimate overflow probabilities in a target node. The same problem has been addressed in [1] and [6], where explicit formulas of the importance function are provided only for very simple networks. We take as a basis the importance function derived in [11] for Jackson networks, in which the function is expressed as a linear combination of the lengths of the queues in front of the network nodes. A first attempt at extending these formulas to non-Markovian networks was made in [10], where the coefficients given by the formulas provided in [11] were multiplied by correction factors adjusted by means of pilot runs. Although good results were obtained for two network topologies studied with exponential interarrival and Erlang service times, the approach was seen to be impracticable when other networks and time distributions were considered in light of the large number of correction factors that had to be adjusted. This paper presents a new approach, which does not require the correction factors to be adjusted. The key point of the approach is the introduction of the concept of effective load. This concept, which for Jackson networks matches the actual load, allows the importance function derived in [11] to be used for non-Markovian networks by substituting in the formulas the effective load of each node for the actual load. This approach solves in a simple and general way this complicated and challenging problem, which has no analytical or numerical solution and which cannot be solved with importance sampling except in the case of very simple networks. The approach is applied to four network topologies: the two topologies considered in [10], a three-queue tandem network and a network with 7 nodes, and two additional ones, a large network with 15 nodes and a network with 2 nodes and very strong feedback. Exponential, hyperexponential, Erlang, Weibull and lognormal distributions are used to model the interarrival and service times.

The paper is organized as follows: Section 2 presents a review of the method; Section 3 provides the proposed importance function; Section 4 describes the tested networks; Section 5 presents the results of the simulation tests; and, finally, Section 6 states the conclusions.

2. Review of RESTART

2.1. Description

RESTART has been described in several papers, e.g., [12–14]. Nevertheless, it is briefly described here to provide a self-contained paper.

Let Ω denote the state space of a process $X(t)$ and A the rare set whose probability is to be estimated. A nested sequence of sets of states C_i, $C_1 \supset C_2 \supset, \ldots, \supset C_M$ is defined, which determines a partition of the state space Ω into the regions $C_i - C_{i+1}$; the higher the value of i, the higher the importance of the region $C_i - C_{i+1}$, that is., the higher the chance of reaching the rare set. These sets are defined by means of a function, $\Phi : \Omega \to \Re$, called the importance function. Thresholds of Φ, T_i $(1 \leqslant i \leqslant M)$, are defined so that each set C_i is associated with $\Phi \geqslant T_i$.

RESTART works as follows: a simulation path, called main trial, is performed in the same way as if it were a crude simulation. Each time the main trial enters set C_1, the entry state is saved and $R_1 - 1$ simulation retrials of level 1 are performed. Each retrial starts with the saved state and finishes when the retrial exits set C_1. When set C_2 is reached in the main trial or in a retrial of level 1, $R_2 - 1$ retrials of level 2 are performed, each one finishing when they leave set C_2, and so on. Note that the oversampling made in the region $C_i - C_{i+1}$ (C_M if $i = M$) is given by the accumulative number of retrials: $r_i = \prod_{j=1}^{i} R_j$.

Fig. 1 illustrates a RESTART simulation with $M = 3$, $R_1 = R_2 = 4$, $R_3 = 3$, in which the chosen importance function Φ also defines set A as $\Phi \geqslant L$. Bold, thin, dashed and dotted lines are used to distinguish the main trial and the retrials of level 1, 2 and 3, respectively.

At the beginning of each retrial it is convenient to reschedule all future events that have been previously scheduled. If this is not done, the correlation between retrials will be high because future events will be shared and, as a consequence, RESTART will be less effective.

The rare set probability, $P = \Pr\{A\}$, may be defined as the probability of the system being in a state of set A at the instant certain events, denoted reference events, occur. For example, in the study presented in this paper, a reference event is the arrival of a customer at the target queue. A reference event in which the system is in a state of the rare set A is referred to as a rare event A. If the rare set A is included in C_M, the estimator of P in a RESTART simulation is $\hat{P} = \frac{N_A}{Nr_M}$, where N_A is

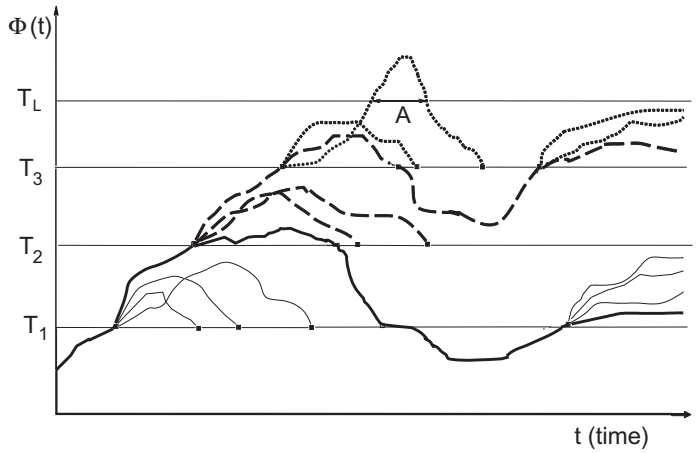

Fig. 1. Simulation with RESTART.

the total number of rare events A that occur in the simulation (in the main trial or in any retrial) and N the number of reference events simulated in the main trial.

2.2. Gain obtained

The gain (also called speedup) obtained with an acceleration method is defined as the ratio of the computational time required for a crude simulation to that required for an accelerated simulation to estimate the same probability with the same width of confidence interval. The gain, G, obtained with RESTART is given by [13]:

$$G = \frac{1}{f_T f_R f_O f_V} \frac{1}{P(-\ln P + 1)^2}. \tag{1}$$

The term $1/P(-\ln P + 1)^2$ can be considered the ideal gain, and factors f_T, f_R, f_O and f_V inefficiency factors that reduce the actual gain with respect to the ideal one.

Formulas of the four factors and criteria for minimizing f_T, f_R and f_O, which are equal to or greater than 1, were provided in [13]. Factors f_T and f_R reflect the inefficiencies due to the non-optimal choice of thresholds or of the number of retrials, respectively. These two factors could easily be made lower than 1.5 in all the cases considered here. Factor f_O reflects the inefficiency due to the computational overheads associated with the implementation of RESTART. This factor affects the computational time but not the number of events that have to be simulated.

Factor f_V is the most critical factor and reflects the inefficiency due to the non-optimal choice of the importance function. Let Ω_i denote the set of possible system states x_i reached when the process enters set C_i. The importance of the state $x_i \in \Omega_i$ is defined as the expected number of rare events A in a retrial of level i starting with the state x_i. Factor f_V (see [13]) is close to 1 if all the states $x_i \in \Omega_i$ are of similar importance, i.e., when the variance of the importance of states x_i for each level i is close to 0. Since f_V increases when these variances increase, the main concern when choosing the importance function must be to reduce these variances.

3. Importance function for queueing networks

This section deals with the problem of obtaining a formula for the importance function in non-Markovian networks for which the rare set is defined as the number of customers in a target node exceeding a predefined threshold. In [11], a formula for the importance function in Jackson networks was obtained. Although the formula was derived using several approximations, it led to small values of factor f_V in most cases. This formula is shown in Section 3.1 and is extended to non-Markovian networks in Section 3.2.

Let us describe first the networks to which the extended formula of the importance function is applied. A queueing network with any finite number of nodes, N, is considered. The buffer space in the queue of each node is assumed to be infinite. Customer arrivals from outside the network are allowed at all the nodes. Interarrival and service time distributions at each node may be Markovian or non-Markovian. After a customer has been served at the lth node, it proceeds to the mth node with probability $p_{l,m}$ or departs from the entire network with probability $p_{l,0}$. The external arrival rate at the lth node is denoted by γ_l, and the total arrival rate (including the arrivals from other nodes) by λ_l. The service rate of the lth node is μ_l and the load is $\rho_l = \lambda_l/\mu_l$. The concepts of effective load, ρ_l^e, and of effective service rate, μ_l^e, are introduced in Section 3.2 for non-Markovian networks. For uniformity, the same notation is used for Markovian networks, but for these networks $\rho_l^e = \rho_l$ and

$\mu_l^e = \mu_l$. The objective of the simulation is to estimate the steady-state probability of the number of customers exceeding a level at a target node.

Let t_g denote the target node, $i^{(1)}$ $(1 \leqslant i \leqslant H_1)$ a node at distance 1 from the target (i.e., a node for which $p_{i^{(1)},t_g} > 0$), and $i^{(2)}$ $(1 \leqslant i \leqslant H_2)$ a node at distance 2 (i.e., a node for which $p_{i^{(2)},t_g} = 0$ and $p_{i^{(2)},j^{(1)}} > 0$ for some value of j). Although some network nodes are at a distance greater than 2 from the target node, they are not taken into account in the formula of the importance function because the dependence of the target node on these nodes is usually very weak.

3.1. Jackson networks

As mentioned above, a formula for the importance function in Jackson networks was derived in [11]. With some changes of notation, the formula can be rewritten as follows:

$$\Phi = Q_{t_g} + \sum_{r=1}^{2}\sum_{i=1}^{H_r} A_{i^{(r)}} Q_{i^{(r)}}, \tag{2}$$

where Q_{t_g} and $Q_{i^{(r)}}$ are the queue lengths of nodes t_g and $i^{(r)}$, respectively, and $A_{i^{(r)}}$ are coefficients which reflect the impact of the queue length $Q_{i^{(r)}}$ on the future evolution of Q_{t_g}. These coefficients are given by:

$$A_{i^{(1)}} = \min\left\{1, \alpha_{i^{(1)}} \frac{\ln(\rho^e_{t_g}/\rho^{i^{(1)}}_{t_g})}{\ln \rho^e_{t_g}}\right\}, \quad A_{i^{(2)}} = \sum_{j=1}^{H_1} \frac{p_{i^{(2)},j^{(1)}}}{\sum_{l=1}^{H_1} p_{i^{(2)},l^{(1)}}} \min\left\{A_{j^{(1)}}, \alpha_{i^{(2)}} \frac{\ln(\rho^e_{t_g}/\rho^{i^{(2)}}_{t_g})}{\ln \rho^e_{t_g}}\right\}. \tag{3}$$

$\rho^{i^{(r)}}_{t_g}$ is the approximate load of the target queue when node $i^{(r)}$ is not empty, and $\alpha_{i^{(r)}}$ is a factor which reflects how the impact of the queue length $Q_{i^{(r)}}$ on the future evolution of Q_{t_g} is affected by the presence of the other queues of the network. $\rho^{i^{(r)}}_{t_g}$ and $\alpha_{i^{(r)}}$ are given by:

$$\rho^{i^{(1)}}_{t_g} = \frac{\lambda_{t_g} + (\mu^e_{i^{(1)}} - \lambda_{i^{(1)}}) p_{i^{(1)},t_g}}{\mu^e_{t_g}}, \tag{4}$$

$$\rho^{i^{(2)}}_{t_g} = \frac{\gamma_{t_g} + \sum_{j=1}^{H_1} \min\{\lambda_{j^{(1)}} + (\mu^e_{i^{(2)}} - \lambda_{i^{(2)}}) p_{i^{(2)},j^{(1)}}, \mu^e_{j^{(1)}}\} p_{j^{(1)},t_g} + \lambda_{t_g} p_{t_g,t_g}}{\mu^e_{t_g}}. \tag{5}$$

$$\alpha_{i^{(1)}} = 1 + \frac{\sum_{l=1,l\neq i}^{H_1}\left(\sum_{j=1}^{H_2}\gamma_{j^{(2)}} p_{j^{(2)},l^{(1)}} + \gamma_{l^{(1)}}\right) p_{l^{(1)},t_g} + \gamma_{t_g}}{\mu^e_{i^{(1)}} p_{i^{(1)},t_g}}, \quad \alpha_{i^{(2)}} = 1 + \frac{\sum_{l=1}^{H_1}\left(\sum_{j=1,j\neq i}^{H_2}\gamma_{j^{(2)}} p_{j^{(2)},l^{(1)}} + \gamma_{l^{(1)}}\right) p_{l^{(1)},t_g} + \gamma_{t_g}}{\mu^e_{i^{(2)}} \sum_{l=1}^{H_1} p_{i^{(2)},l^{(1)}} p_{l^{(1)},t_g}}. \tag{6}$$

In spite of the apparent complexity of these formulas, they are easy to apply because all their terms represent parameters of the system. For the particular case of a three-queue tandem network with the third node as target, formulas (2)–(6) become:

$$\Phi = Q_{t_g} + \sum_{r=1}^{2} A_{1^{(r)}} Q_{1^{(r)}} \quad \text{with}: \quad A_{1^{(1)}} = \min\left\{1, \frac{\ln(\rho^e_{1^{(1)}})}{\ln \rho^e_{t_g}}\right\}, \quad A_{1^{(2)}} = \min\left\{A_{1^{(1)}}, \frac{\ln(\rho^e_{1^{(2)}})}{\ln \rho^e_{t_g}}\right\}. \tag{7}$$

where nodes $1^{(2)}$, $1^{(1)}$ and t_g are the first, second and third network nodes respectively.

In [11], the above formulas were not only used to evaluate the importance of the states outside the rare set, but also of the states within. As a consequence, the rare set, A, was not always included in the last importance region, C_M, leading to lower efficiency in some cases. Thus, in the present paper, we have made $\Phi = T_M$ in all the states within the rare set to ensure that $A \subset C_M$ and to improve efficiency.

3.2. Non-Markovian networks

As can be seen from the formulas of Section 3.1, the proposed importance function is a linear combination of the queue lengths of the target node and of the nodes that are at a distance 1 or 2 from it. Each coefficient of the formula represents the weight that must be given to the corresponding queue length when the importance of a given state is evaluated. Note that the coefficient corresponding to a node is a function of the load of the node: if the load is 1 the coefficient is 0 and, in general, the higher the load, the lower the value of the coefficient. This can be intuitively explained as follows: when the load of a node is high, its queue length is usually high, and thus a high value of the queue length does not imply a system state of high importance. Hence, a low weight should be given to the queue length of that node in the evaluation of the importance. In contrast, a high weight should be given to the queue length of a node for which the queue length is usually low. We can say that the coefficient of a node is a function of its load because the load is an indicator of its queue length distribution. Indeed, the queue length distribution, $\Pr\{Q \geqslant n\}$, and the load, ρ, of a node of a Jackson network are related by: $\rho = [\Pr\{Q \geqslant n\}]^{1/n}$.

In a non-Markovian network, the nth root of $\Pr\{Q \geqslant n\}$ is not the load of the node, but we can define an effective load, ρ^e, to represent this nth root:

$$\rho^e = \lim_{n\to\infty}[\Pr\{Q \geq n\}]^{1/n}. \tag{8}$$

Based on this reasoning and as an approximation, we propose applying formulas (2)–(7), as in the case of Jackson networks, but substituting the effective load of the lth node, ρ_l^e, for the actual load, ρ_l (and the effective service rate, μ_l^e, given by $\mu_l^e = \lambda_l/\rho_l^e$, for its actual service rate, μ_l). The approximations made in the formula of the importance function affect the efficiency of the method but do not affect the correctness of the estimates, given that the estimator of the rare set probability is unbiased for any importance function, as proven in [13]. The goodness of the proposed approximation is supported by the efficient application of the method shown in Section 5.

In practice, formula (8) must be evaluated for a finite value of n. The probability $\Pr\{Q \geqslant n\}$ may be estimated in a pilot run made using crude simulation. The effective load, ρ^e, may be evaluated using the expression $[\Pr\{Q \geqslant n\}/\Pr\{Q \geqslant n/2\}]^{2/n}$, which has the same limit as $[\Pr\{Q \geqslant n\}]^{1/n}$ when $n \to \infty$, but approaches this limit faster, allowing a good approximation with a lower value of n. The value of n used in this paper is 10 for most nodes and smaller for some low loaded nodes.

4. Description of the test cases

We conducted several simulation experiments on networks with different topologies and loads, considering several interarrival and service time distributions. The tested topologies are presented in Section 4.1 and the interarrival and service time distributions in Section 4.2. In all the examples, the maximum value of L satisfying $\Pr\{Q_{t_g} \geq L\} \geq 10^{-15}$ was chosen.

4.1. Tested topologies

4.1.1. Two-node network with strong feedback

We consider a network with two nodes, with an external arrival rate at each node equal to 1. Customers departing any of the two nodes join the other node with a probability of 0.8 or leave the network with a probability of 0.2. The node loads are $\rho_1 = 0.67$, $\rho_2 = 0.33$. The 2nd node is taken as target node. In spite of its simplicity, this example involves a particular difficulty due to the strong feedback between the two nodes and to the fact that the target node is the least loaded.

4.1.2. Three-queue tandem network

Customers arrive at the first queue of this network at an arrival rate equal to 1, then go to the second queue, then to the third queue and finally leave the network. Three load cases, corresponding to three different bottleneck nodes, are considered: in case a, $\rho_1 = 0.67$, $\rho_2 = 0.50$, $\rho_3 = 0.33$; in case b, $\rho_1 = 0.33$, $\rho_2 = 0.50$, $\rho_3 = 0.33$; and in case c, $\rho_1 = 0.20$, $\rho_2 = 0.25$, $\rho_3 = 0.33$.

The third node is taken as target node. The difficulty of estimating $\Pr\{Q_{t_g} \geq L\}$ in a two-queue tandem network when the first node is the bottleneck is well known (see [14] and references therein). The difficulty may be similar in case b, where $\rho_2 > \rho_{t_g}$, and much greater in case a, where $\rho_1 > \rho_2 > \rho_{t_g}$.

4.1.3. Network with 7 nodes

Consider a network with three sets of nodes: set 0 with one node (node 01), set 1 with two nodes (11 and 12) and set 2 with four nodes (21–24). Customers from outside the network arrive at each node at a rate equal to 1. After being served at each node, a customer leaves the network with a probability of 0.2 or goes to the same or to another node in accordance with the transition matrix shown in Table 1.

Three load cases, with different bottleneck nodes, are considered: in case a, $\rho_{01} = 0.33$, $\rho_{1i} = 0.41$, $\rho_{2j} = 0.50$; in case b, $\rho_{01} = 0.33$, $\rho_{1i} = 0.41$, $\rho_{2j} = 0.32$; in case c, $\rho_{01} = 0.33$, $\rho_{1i} = 0.30$, $\rho_{2j} = 0.28$, with $i = 1, 2$ and $j = 1,\ldots,4$ in all three cases. Node 01 is taken as target node.

Table 1
Transition matrix of the 7-node network.

	01	11	12	21	22	23	24
01	**0.2**	0.1	0.1	0.1	0.1	0.1	0.1
11	0.3	**0**	**0.1**	0.1	0.1	0.1	0.1
12	0.3	**0.1**	**0**	0.1	0.1	0.1	0.1
21	0	0.2	0.2	**0.1**	**0.1**	**0.1**	**0.1**
22	0	0.2	0.2	**0.1**	**0.1**	**0.1**	**0.1**
23	0	0.2	0.2	**0.1**	**0.1**	**0.1**	**0.1**
24	0	0.2	0.2	**0.1**	**0.1**	**0.1**	**0.1**

4.1.4. Network with 15 nodes

Consider a network with 4 sets of nodes: set 0 with 3 nodes, denoted by $0i$ ($i = 1,\ldots,3$) and each of the sets 1, 2 and 3 with 4 nodes, denoted by $1i$, $2i$ and $3i$, ($i = 1,\ldots,4$), respectively.

Customers from outside the network arrive at each node at a rate equal to 1. After being served at each node, a customer leaves the network with a probability of 0.2 or goes to the same or to another node in accordance with the transition matrix shown in Table 2.

A service rate equal to 10 was chosen for all the nodes. Node 03, which is one of the three least loaded nodes, was taken as target node. Note that nodes of set 3 are at a distance greater than 2 from the target node, and thus they are not included in the formulas of the importance function. Therefore, this example is useful for checking the validity of this approximation.

4.2. Tested interarrival and service time distributions

In each test run the same interarrival time distribution for external arrivals has been assumed in all the network nodes (except the mean values, which may be different at each node). The same has been assumed for the service time distributions. First, we tested three distributions: exponential, hyperexponential (with $\eta_1 = 6\eta_2$ and $p_1 = p_2 = 0.5$) and Erlang-3, with coefficients of variation of 1, 1.42 and 0.58, respectively. Nine interarrival/service combinations were tested: each of the three interarrival time distributions with each of the three service time distributions. The objective of these tests was to check whether the proposed approach is valid for different values of the coefficient of variation both in the interarrival and in the service time distributions. Additionally, in order to check whether the approach is valid for other distributions, we tested the combination of Weibull interarrival and lognormal service time distributions. The shape parameters of these two distributions were 2 and 0.5, respectively.

Each run will be identified by the abbreviation of the interarrival time distribution followed by the abbreviation of the service time distribution. The exponential–exponential (Exp–Exp) case corresponds to a Jackson network, which was already treated in [11], but the results are repeated here for comparison purposes. The Exp–Erl case was treated in [10] for the networks of 3 and 7 nodes, but the approach followed was different, as explained in Section 1.

5. Results of the tests

Each one of the 8 combinations of networks and loads was simulated with each of the 10 combinations of interarrival and service time distributions. The effective loads were obtained from a previous crude simulation. Thresholds were set for every integer value of Φ. Using a tentative number of retrials, a pilot run was made to estimate the number of retrials, as indicated in [13]. The confidence interval is evaluated using the independent replication method. The simulation finishes when the relative error is not greater than 0.1, i.e., when the half width of the 95% confidence interval is not greater than a 10% of the estimate. Three simulation runs were made for each case and only the results of the run with intermediate computational times were taken into account. All the experiments were run on a standard desktop computer with a 3.20 GHz Intel Core i7 CPU and 8 GB RAM.

Section 5.1 presents the detailed results obtained for the Hyp–Erl cases and Section 5.2 summarizes the results obtained for all the cases. In order to facilitate effective and efficient model sharing, all the details needed to reproduce the simulations have been written in the supplementary material, see Appendix A.

Table 2
Transition matrix of the 15-node network.

	01	02	03	11	12	13	14	21	22	23	24	31	32	33	34
01	**0.1**	**0.1**	**0**	0.1	0.1	0	0	0.1	0.1	0	0	0.1	0.1	0	0
02	**0**	**0.1**	**0.1**	0	0.1	0.1	0	0	0.1	0.1	0	0	0.1	0.1	0
03	**0.1**	**0**	**0.1**	0	0	0.1	0.1	0	0	0.1	0.1	0	0	0.1	0.1
11	0.1	0.1	0.1	**0**	**0.1**	**0**	**0**	0.1	0.1	0	0	0.1	0.1	0	0
12	0.1	0.1	0.1	**0**	**0**	**0.1**	**0**	0	0.1	0.1	0	0	0.1	0.1	0
13	0.1	0.1	0.1	**0**	**0**	**0**	**0.1**	0	0	0.1	0.1	0	0	0.1	0.1
14	0.1	0.1	0.1	**0.1**	**0**	**0**	**0**	0.1	0	0	0.1	0.1	0	0	0.1
21	0	0	0	0.1	0.1	0.1	0.1	**0**	**0.1**	**0.1**	**0**	0.1	0.1	0	0
22	0	0	0	0.1	0.1	0.1	0.1	**0**	**0**	**0.1**	**0.1**	0	0.1	0.1	0
23	0	0	0	0.1	0.1	0.1	0.1	**0.1**	**0**	**0**	**0.1**	0	0	0.1	0.1
24	0	0	0	0.1	0.1	0.1	0.1	**0.1**	**0.1**	**0**	**0**	0.1	0	0	0.1
31	0	0	0	0	0	0	0	0.1	0.1	0.1	0.1	**0.1**	**0.1**	**0.1**	**0.1**
32	0	0	0	0	0	0	0	0.1	0.1	0.1	0.1	**0.1**	**0.1**	**0.1**	**0.1**
33	0	0	0	0	0	0	0	0.1	0.1	0.1	0.1	**0.1**	**0.1**	**0.1**	**0.1**
34	0	0	0	0	0	0	0	0.1	0.1	0.1	0.1	**0.1**	**0.1**	**0.1**	**0.1**

5.1. Hyperexponential interarrival and Erlang service time distributions

5.1.1. Two-node network with strong feedback

As was said in Section 4.1.1, the actual loads of the first and second nodes are 0.67 and 0.33, respectively. The effective loads obtained were 0.75 for the first node (node $1^{(1)}$) and 0.22 for the second node (target node). The importance function is $\Phi = Q_{t_g} + A_{1^{(1)}} Q_{1^{(1)}}$, where $A_{1^{(1)}}$ is equal to 0.19 (much smaller than the value 0.36 obtained with the actual loads).

The value of L defining the rare set was 23, and the resulting rare set probability was $\Pr\{A\} = 2.2 \times 10^{-15}$. The number of reference events (arrivals of customers at the target queue) that had to be simulated was 4.5 million, and the required computational time 0.7 min (without including the time of the pilot runs, which is usually small).

To evaluate the gain in events (i.e., the gain measured in terms of the number of events that have to be simulated) or the gain in time with respect to a crude simulation, the number of events and the computational time of the crude simulations were estimated by extrapolating the measured values for higher probabilities. A gain in events equal to 6.1×10^{10} and a gain in time equal to 4.3×10^{9} were obtained. The smaller value of the gain in time is due to the inefficiency factor f_O which, as explained in Section 2.2, affects the required computational time but not the number of events that have to be simulated.

Considering formula (1), factor f_V, which indicates how appropriate the chosen importance function is, was estimated from the gain in events, as indicated in [14]. The value obtained was $f_v = 6.1$, which is similar to the value obtained in the Exp–Exp case, $f_v = 5.9$. Deriving the coefficient $A_{1^{(1)}}$ from the effective loads was crucial for obtaining this moderate value of f_V. When the value of $A_{1^{(1)}}$ was derived from the actual loads it was not possible to obtain accurate results with a reasonable computational time.

5.1.2. Three-queue tandem network

In this network, except in case c, the effective loads are also significantly different from the actual loads. For example, in case a, the actual loads are $\rho_1 = 0.67$, $\rho_2 = 0.50$ and $\rho_3 = 0.33$, while the effective loads are $\rho_1^e = 0.72$, $\rho_2^e = 0.29$ and $\rho_3^e = 0.08$. Thus, the coefficients of the importance function derived from the effective loads are also very different from those derived from the actual loads.

The required computational times were 5.5, 2.1 and 0.3 min for cases a, b and c, respectively, while the values of factor f_V were 70, 19 and 2.8. These moderate values of f_V show that the choice of the importance function was appropriate. The worst result was obtained when $\rho_1 > \rho_2 > \rho_{t_g}$ (case a), but, even in this critical case, the computational time was moderate.

In contrast to the good results obtained with the coefficients derived from the effective loads, it was not possible, except in case c, to obtain accurate results within a reasonable time using the coefficients derived from the actual loads.

5.1.3. Network with 7 nodes

In this network, the values of the effective loads are very close to the respective values of the actual loads. For example, in case a, the actual loads are $\rho_{01} = 0.33$, $\rho_{1i} = 0.41 (i = 1, 2)$ and $\rho_{2j} = 0.50 (j = 1, \ldots, 4)$, and the effective loads 0.33, 0.39 and 0.51 respectively. As a consequence, the coefficients of the importance function are very similar using the actual loads and the effective loads.

The results obtained were very good in all three considered cases. The computational times were 0.6, 0.4 and 0.4 min in cases a, b and c, respectively and factor f_V took the values of 2.1, 1.4 and 1.1, respectively. These values are very close to 1, meaning that the importance function is very close to the optimal in all the cases. As could be expected, the worst results were obtained when $\rho_{2j} > \rho_{1i} > \rho_{t_g}$ (case a), while the best results were obtained when the bottleneck was the target queue (case c). The results are better than those obtained in the previous networks because the target node is not visited by many of those customers that visit the other network nodes. Thus, the dependence of the target queue on the length of the other queues is weaker, and the achieved efficiency is greater. Unlike importance sampling, RESTART may improve its efficiency with the complexity of the systems.

The results obtained using the coefficients derived from the actual loads are also good ($f_V = 3.3$ in case a, 1.4 in case b and 1.8 in case c). This is not unexpected given that for this network the coefficients obtained with the two approaches have similar values.

5.1.4. Network with 15 nodes

As in the network with 7 nodes, the two sets of loads take similar values and, as a consequence, the two corresponding sets of coefficients of the importance function are also similar.

The required computational time was 5.0 min and the value of factor f_V was 4.0. This moderate value indicates that large networks are not a problem for RESTART and that the nodes that are at a distance greater than 2 from the target node do not have to be taken into account in the formula of the importance function.

Similar results ($f_V = 3.4$) were obtained using the coefficients derived from the actual loads, as could be expected given that the two sets of coefficients have similar values.

A sensitivity study was made for the eight Hyp–Erl cases described above, testing importance functions with all the coefficients (except that of the target queue) that were within 10% (above or below) those obtained using the effective loads. In six of the eight cases, the best computational times were obtained with the coefficients obtained using the effective loads, while, for the 2- and 15-node networks, slightly smaller times were obtained with the coefficients that were 10% lower.

Table 3
Computational time (min) for estimating a probability of 10^{-15}. Relative error $\leqslant 0.1$.

Interarrival law		Exp			Erl			Hyp			Wei	Mean
Service law		Exp	Erl	Hyp	Exp	Erl	Hyp	Exp	Erl	Hyp	Log	
2 nodes		0.5	0.5	3.0	1.9	5.8	1.6	1.1	0.7	3.7	6.4	**2.3**
3 nodes	Case a	1.1	6.1	3.9	2.2	7.1	5.7	3.3	5.5	2.8	42.9	**8.1**
	Case b	0.3	1.8	1.4	0.5	3.5	0.9	0.7	2.1	1.0	33.1	**4.5**
	Case c	0.1	0.3	0.1	0.1	0.4	0.3	0.1	0.3	0.2	8.5	**1.0**
7 nodes	Case a	0.4	0.5	0.4	0.9	1.3	1.0	0.7	0.8	4.0	1.2	**1.1**
	Case b	0.2	0.3	0.3	0.4	0.6	0.4	0.2	0.6	1.7	0.7	**0.5**
	Case c	0.2	0.3	0.5	0.2	0.4	0.4	0.3	0.6	0.3	0.5	**0.4**
15 nodes		1.1	1.5	2.1	1.0	1.2	2.7	2.4	4.2	4.3	5.7	**2.6**
Mean		**0.5**	**1.4**	**1.5**	**0.9**	**2.5**	**1.6**	**1.1**	**1.9**	**2.3**	**12.4**	**2.5**

5.2. Other interarrival and service time distributions

As mentioned above, each one of the 8 combinations of networks and loads was simulated with each of the 10 combinations of interarrival and service time distributions. The values of L were chosen as indicated in Section 4 to give rare set probabilities in the order of 10^{-15}.

In the 2- and 3-node networks the coefficients of the importance function were obtained with the effective loads. They were generally quite different from those obtained from the actual loads and led to much more efficient applications. In the 7- and 15-node networks, as the coefficients and the efficiency obtained for the Hyp–Erl case were similar using either the effective load or the actual load, it was considered not worth evaluating the effective loads. Therefore, the results shown in this section for these networks correspond to coefficients obtained from the actual loads. As a general rule, we recommend first obtaining the coefficients of the importance function from the actual loads, at least in the case of complex networks. Only if accurate estimations cannot be obtained within a reasonable time with these coefficients, will it be necessary to evaluate the effective loads and to use the coefficients obtained from them.

Table 3 shows the required computational times (excluding the time of the pilot runs, which is usually small). The time is small in all cases: 42.9 min in the worst case and 2.5 min on average for all 80 cases. This underlines the applicability of RESTART as an efficient acceleration method and the suitability of the proposed importance function. It was observed that the computational times were, in general, smaller for the 7- and 15-node networks than for the 2- and 3-node networks, contrary to what we expected based on our experience with crude simulations. However, it confirms what was said in Section 5.1: RESTART may improve with the complexity of the systems due to the weaker dependence of the target queue on the length of the other queues. A comparison of the computational times in the three load cases of the 3- and 7-node networks confirms the greater difficulty that arises when the target node is the least loaded (case a). The greater computational times required in the Wei–Log cases are partially due to the greater computational times needed to generate random numbers from these distributions.

As mentioned in Section 1, the estimator used in RESTART simulations is unbiased regardless of the importance function chosen. This is corroborated here in the Exp–Exp cases, for which analytical results are available. For example, in the two-node network, the rare set probability obtained from the simulation was $\Pr\{Q_{t_g} \geq 31\} = 1.53 \cdot 10^{-15}$. As the load of the target node is 1/3, the analytical result is $(1/3)^{31} = 1.62 \times 10^{-15}$, which is within the confidence interval $(1.38 \times 10^{-15}, 1.68 \times 10^{-15})$. The same occurs in the other Exp–Exp cases; for example, in the network with 15 nodes, $L = 32$ and $\rho_{t_g} = 0.341$, the simulation and analytical results for $\Pr\{Q_{t_g} \geq 32\}$ are 1.21×10^{-15} and $0.341^{32} = 1.12 \times 10^{-15}$, respectively.

6. Conclusions

The choice of importance function is the most critical feature when RESTART is applied to multi-dimensional systems. This paper has focused on finding effective importance functions for estimating the overflow probability of a target queue in non-Markovian queueing networks. A simple and innovative concept, the effective load of a node, defined as the actual load of a node of a Jackson network with a similar queue length distribution, is introduced. Despite its simplicity, this concept has allowed this challenging problem to be solved in a general and easy way by substituting the effective loads for the actual loads in the formulas of the importance functions previously derived for Jackson networks.

The formulas have been tested with four network topologies, ranging from two simple ones, a two-queue network with strong feedback and a three-queue tandem network, to two complex ones with 7 and 15 nodes and multiple feedbacks.

Exponential, hyperexponential, Erlang, Weibull and lognormal distributions have been used to model the interarrival and service times. Ten combinations of interarrival/service time distributions were tested for each topology and for each considered load case, resulting in a total of 80 case studies. In all of them, overflow probabilities of around 10^{-15}, which is lower than needed in practical problems, were accurately estimated with short computational times (2.5 min on average for all 80

cases). The results obtained show that the complexity of the network does not represent a barrier for the effective application of the method, but, on the contrary, the efficiency often improves with the complexity. The wide variety of cases tested and the difficulty inherent in most of them (high dependence between queue lengths, strong feedback and/or network complexity) suggests that the method may be effective for many other Jackson and non-Jackson queueing networks. It should be noted that importance sampling has never been applied to complex networks such as those shown in this paper, an application which, according to the current literature (see, e.g., [3]), would not seem feasible. As no other acceleration method is able to solve this problem and, except for exponential interarrival and service time distributions, no analytical results are available, this paper must be considered a pioneer in that it is the first to evaluate rare event probabilities for these networks.

The concept of effective load introduced in the paper is a promising concept with potential applications to other network features, such as burst arrivals, task priorities or several customer types, each type with different service times and transition probabilities. Future research will focus on extending the concept of effective load to cover these other features, the ultimate goal being to obtain a method that can be easily implemented in any queueing network simulation tool, such as JSIM [2].

Acknowledgments

This research was partially supported by the Comunidad of Madrid, grant Riesgos CM (P2009/ESP-1685) and by CYCYT, grant Evaluación y Gestión del Riesgo en MCDM.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.simpat.2013.05.012.

References

[1] M. Amrein, H.R. Kunsch, A variant of importance splitting for rare event estimation: fixed number of successes, ACM Trans. Model. Comput. Simul. 21 (2) (2011). Article 13.
[2] M. Bertoli, G. Casale, G. Serazzi, JMT: performance engineering tools for system modeling, ACM SIGMETRICS Perform. Eval. Rev. 36 (4) (2009) 10–15.
[3] Z.I. Botev, D.P. Kroese, An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting, Meth. Comput. Appl. Probab. 10 (4) (2008) 471–505.
[4] M.J.J. Garvels, A combined splitting-cross entropy method for rare-event probability estimation of queueing networks, Ann. Oper. Res. 189 (2011) 167–185.
[5] P.E. Heegard, W. Sandmann, Importance sampling simulations of phase-type queues, in: Proceedings of the 2009 Winter Simulation Conference, Austin, 2009, pp. 1136–1145.
[6] S. Juneja, V. Nicola, Efficient simulation of buffer overflow probabilities in Jackson networks with feedback, ACM Trans. Model. Comput. Simul. 15 (4) (2005) 281–315.
[7] H. Kahn, T.E. Harris, Estimation of particle transmission by random sampling, National Bureau of Standards Appl. Math. Series 12 (1951) 27–30.
[8] P. L'Ecuyer, V. Demers, B. Tuffin, Rare events, Rare events, splitting and quasi-Monte Carlo, ACM Trans. Model. Comput. Simul. 17 (2) (2007). Article 9.
[9] G. Rubino, B. Tuffin (Eds.), Rare Event Simulation using Monte Carlo Methods, John Wiley, Chichester, 2009.
[10] J. Villén-Altamirano, RESTART simulation of networks of queues with Erlang service times, in: Proceedings of the 2009 Winter Simulation Conference, Austin, 2009, pp. 1146–1154.
[11] J. Villén-Altamirano, Importance function for RESTART simulation of general Jackson networks, Eur. J. Oper. Res. 203 (1) (2010) 156–165.
[12] M. Villén-Altamirano, J. Villén-Altamirano, RESTART: a method for accelerating rare event simulations, in: Proceedings of the International Teletraffic Congress, ITC 13, Queueing, Performance and Control in ATM North Holland, Amsterdam, 1991, pp. 71–76.
[13] M. Villén-Altamirano, J. Villén-Altamirano, Analysis of RESTART simulation: theoretical basis and sensitivity study, Eur. Trans. Telecomm. 13 (4) (2002) 373–386.
[14] M. Villén-Altamirano, J. Villén-Altamirano, On the efficiency of RESTART for multidimensional state systems, ACM Trans. Model. Comput. Simul. 16 (3) (2006) 251–279.