Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds.

RESTART SIMULATION OF NETWORKS OF QUEUES WITH ERLANG SERVICE TIMES

José Villén-Altamirano

Dept. of Applied Mathematics
Polytechnic University of Madrid
Calle Arboleda s/n, 28031 Madrid, Spain

ABSTRACT

RESTART is an accelerated simulation technique that allows the evaluation of low probabilities. 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 the rare event is higher. These regions are defined by means of a function of the system state called the importance function. Guidelines for obtaining suitable importance functions and formulas for the importance function of general Jackson networks were provided in previous papers. In this paper, we study networks with Erlang service times and with the rare set defined as the number of customers in a target node exceeding a predefined threshold. The coefficients of the importance functions used here are the same as those obtained with the formula for Jackson networks but multiplied by a constant obtained heuristically. Low probabilities are accurately estimated for different network topologies within short computational time.

1 INTRODUCTION

The study of critical events that occur very infrequently is of interest in many areas. The performance requirements of broadband communication networks and ultra reliable systems are often expressed in terms of events with very low probability. Developing good estimates of quality of service provided by a network, requires studying scenarios that may occur rarely during the life of the network. Analytical or numerical evaluation of low probabilities is only possible for a very restricted class of systems due to the model assumptions that are needed. Although simulation is an effective means of studying such systems, acceleration methods are necessary because crude simulations require prohibitively long execution times for the accurate estimation of very low probabilities.

Importance sampling is a well known technique in rare event simulation, see e.g., Rubino and Tuffin (2009). The basic idea behind this approach is to alter the probability measure governing events so that the formerly rare event occurs more often. One drawback of this technique is the difficulty of selecting an appropriate change of measure since it depends on the system being simulated. Most research has focused on finding good heuristics for particular types of models. Dupuis and Wang (2008) deal with the construction of asymptotically optimal importance sampling schemes for queueing networks based on subsolutions to an associated partial differential equation. Importance sampling has difficulties to deal with large systems and/or systems with strong feedback.

Another known method is RESTART (REpetitive Simulation Trials After Reaching Thresholds), which is based on a completely different idea. In this method a more frequent occurrence of a formerly rare event 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. The retrials are killed if they go to a region with lower importance than the region where they were generated. Villén-Altamirano and Villén-Altamirano (1991) coined the name RESTART and made a theoretical analysis that yields the variance of the estimator and the gain obtained with one threshold. A rigorous analysis of multiple thresholds was made in Villén-Altamirano and Villén-Altamirano (2002), where optimal values for thresholds and the number of retrials that maximize the gain were derived.

A precedent of much more limited scope is the splitting technique described in Kahn and Harris (1951). See also chapter 3 of Rubino and Tuffin (2009) and the references therein. In this method retrials are also performed but only the first time the process enters importance regions and the retrials are not killed when they go to lower importance regions. As a huge computational time is wasted simulating such unpromising trials this method is only valid for simulations made by means of short replicas, e.g., regenerative simulations of very simple systems, or short transient simulations. But even for simple cases Dean

978-1-4244-5771-7/09/$26.00 ©2009 IEEE

and Dupuis (2009) showed that the computational time of their application of RESTART (they called it RESTART/DPR) ranges from 14% to 4% of the corresponding time required with splitting for the same simple systems. DPR (Direct Probability Redistribution) was introduced by Haraszti and Towsend (1999) and the authors claimed that it was a generalization of RESTART. The only true generalization is that DPR does not require a nested sequence of importance regions, but this possibility has not been ever applied because it would lead to an inefficient algorithm. The other claimed generalizations, as the possibility of "jumping thresholds", or the possibility of reaching the rare set from several regions, were previously contemplated with RESTART.

The application of RESTART or Splitting to particular models requires the choice of a suitable importance function. The problem of finding the optimal importance function has been compared by Garvels et al. (2002) with the problem of finding a good change of measure in importance sampling, because in both cases "knowledge about the behaviour of the system leading to the rare event is necessary". Glasserman et al. (1999) stated that "splitting ultimately relies on a detailed understanding of a process's rare event asymptotic, much as importance sampling does". They suggested that it may be difficult to use splitting or RESTART for systems with multidimensional state space. Villén-Altamirano and Villén-Altamirano (2006) showed that it was possible to obtain heuristically effective importance functions for multidimensional systems without using large deviation theory. Dean and Dupuis (2009) studied the construction of asymptotically optimal RESTART/ DPR algorithms based on subsolutions. Unlike with importance sampling, which requires *classical* sense subsolutions, RESTART requires subsolutions only in the viscosity sense. Villén-Altamirano (2009) obtained a formula of the importance function for Jackson networks combining heuristic arguments with analytical results. Different types of networks with different loads of the nodes were simulated and very low overflow probabilities were accurately estimated within reasonable computational work.

It is interesting to study whether the importance function derived for Jackson networks would be fit for other networks. In this paper we will study networks with Poisson arrivals and Erlang service times. We will compare the results obtained with the importance function derived in Villén-Altamirano (2009) with the results obtained with other importance functions whose coefficients are the same as those obtained with the formula for Jackson networks but multiplied by a correction factor obtained heuristically.

The paper is organized as follows: Section 2 presents a review of the method, Section 3 describes the system under study and Section 4 provides the simulation study.

2 DESCRIPTION OF RESTART

RESTART has been described in several papers, e.g., Villén-Altamirano and Villén-Altamirano (2002, 2006). Nevertheless it is briefly described here.

Let Ω denote the state space of a process $X(t)$ and A a rare subset of the state space whose probability must 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 regions $C_i - C_{i+1}$; the higher the value of i, the higher the importance of the region $C_i - C_{i+1}$. These sets are defined by means of a function $\Phi : \Omega \to \Re$, called the importance function. Thresholds T_i $(1 \leq i \leq M)$ of Φ are defined so that each set C_i is associated with $\Phi \geq T_i$.

RESTART works as follows: each time the process enters a set C_i, the system state is saved and R_i trials of level i are performed. Each trial of level i is a simulation path that starts with the saved state and finishes (except the last one) when it leaves set C_i. The last trial, which continues after leaving set C_i, potentially leads to new sets of trials of level i if set C_i is reached again. A set C_{i+1} may be reached in a trial of level i and an analogous process is set in motion: R_{i+1} trials of level i+1 are performed, and so on. In the case that the process crosses more than one threshold in a time step, the retrials of all the crossed thresholds must be made. If, for instance, a set C_{i+1} is reached in a trial of level i-1, as set C_i is also reached, it is considered that R_i trials of level i are performed and that all of them reach threshold $i + 1$. Thus, $R_i \cdot R_{i+1}$ trials of level i+1 are performed, R_i of them finish when threshold i is crossed down, and the other ones when threshold $i + 1$ is crossed down. This procedure for "jumping thresholds" was implicit in the description of RESTART, see Villén-Altamirano et al. (1994). In fact, it was implemented in the simulation tool ASTRO (Advanced Simulation Tool with RESTART Optimization) described in that paper. Some more notations:

- $P = \Pr\{A\}$ is the probability of the system being in a state of the set A at the instant of occurrence of certain events denoted reference events (e.g., customer arrivals);
- $C_{M+1} = A$;

- $P_{h/i}$ $(0 \le i \le h \le M+1)$: probability of the set C_h at a reference event, knowing that the system is in a state of the set C_i at that reference event. For $h \le M$, as $C_h \subset C_i$, $P_{h/i} = \Pr\{C_h\}/\Pr\{C_i\}$;
- $r_i = \prod_{j=1}^{i} R_j,\ 1 \le i \le M$
- Ω_i $(1 \le i \le M)$: set of possible system states x_i, when the process enters set C_i;
- P^*_{A/x_i} $(1 \le i \le M)$: importance of state x_i, defined as the expected number of events A in a trial of level i starting with that system state;
- $P^*_{A/i}$ $(1 \le i \le M)$: expected importance when the process enters set C_i:

$$P^*_{A/i} = E\left[P^*_{A/X_i}\right] = \int_{\Omega_i} P^*_{A/x_i}\, dF(x_i) \ ,$$

where $F(x_i)$ is the distribution function of X_i;

- $V\left(P^*_{A/X_i}\right)$ $\left(1 \le i \le M\right)$: variance of the importance when the process enters set C_i:

$$V\left(P^*_{A/X_i}\right) = \int_{\Omega_i} \left(P^*_{A/x_i}\right)^2 dF(x_i) - (P^*_{A/i})^2 .$$

If the rare set A is included in C_M, the estimator of the probability of A in a RESTART simulation is: $\hat{P} = \dfrac{N_A}{Nr_M}$, where N_A is the number of events A that occur in the simulation and N the number of reference events simulated without counting the retrials. Otherwise, the estimator is: $\hat{P} = \sum_{i=1}^{M} \dfrac{N_{Ai}}{Nr_i}$, where N_{Ai} is the number of reference events A that occur in the region $C_i - C_{i+1}$ (C_M if $i = M$). The formulas given below in this section have been derived for the first case.

The gain (also called speedup) G obtained with RESTART can be defined as the ratio of the computer cost times the variance of the crude simulation estimator to the computer cost times the variance of the RESTART estimator, see Villén-Altamirano and Villén-Altamirano (2002). In that paper, it is proved that G is given by:

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

The factors f_V, f_O, f_R and f_T, all of them equal to or greater than 1 (with the exception of f_V which may be smaller than 1 in some cases), can be considered inefficiency factors that reduce the actual gain with respect to the term $1\Big/\left(P\left(-\ln P + 1\right)^2\right)$. This term can be considered the ideal gain because it is the maximum gain that can be obtained (except in the cases where $f_V < 1$). Each factor reflects:

- f_V: inefficiency due to the non-optimal choice of the importance function.
- f_O: inefficiency due to the computer overhead of RESTART.
- f_R: inefficiency due to the non-optimal choice of the number of retrials.
- f_T: inefficiency due to the non-optimal choice of the thresholds.

In Villén-Altamirano and Villén-Altamirano (2002) criteria for minimizing the factors f_O, f_R and f_T were given. A value of the factor f_R equal to 1 is achieved if the accumulated number of trials is chosen according to:

$$r_i = \frac{1}{\sqrt{P_{i+1/1} \cdot P_{i/0}}} \ , \ i = 1, \ldots, M . \tag{2}$$

As the number of trials R_i must be an integer number, a value of f_R very close to 1 can be achieved with the following rounding algorithm: R_1 is made equal to r_1 rounded to an integer number, $R_2 = r_2 / R_1$ rounded to an integer number, … ,

$R_i = r_i / \left(R_1 \cdot \ldots \cdot R_{i-1} \right)$ rounded to an integer number. An alternative for obtaining integer number of retrials, proposed in Haraszty and Towsend (1999), is to randomize the number of trials R_i for achieving an expected number of R_i equal to r_i / r_{i-1}.

The factor f_T is minimized by choosing very close thresholds, i.e., $P_{i+1/i}$ close to one. An upper bound of f_T is given by, see Villén-Altamirano and Villén-Altamirano (2002):

$$f_T \le \frac{1/P_{\min} + P_{\min} - 2}{\left(\ln P_{\min} \right)^2} \quad \text{where} \quad P_{\min} = \underset{0 \le i \le M}{Min} \left(P_{i+1/i} \right). \tag{3}$$

The factor f_O affects to the computational time but not to the number of events to be simulated. This factor usually takes moderate values.

In Villén-Altamirano and Villén-Altamirano (2006) the factor f_V was analyzed and guidelines for choosing the importance function were provided. One of the guidelines for reducing f_V is to reschedule the scheduled events at the beginning of each trial. An upper bound of f_V was also given in the paper: $f_V \le \underset{1 \le i \le M+1}{Max} \left(1 + V\left(P^*_{A/X_i} \right) \Big/ \left(P^*_{A/i} \right)^2 \right)$. Thus, the main concern is to minimize $V\left(P^*_{A/X_i} \right)$. It can be achieved by a proper choice of the importance function.

In Villén-Altamirano (2009) a formula of the importance function valid for estimating overflow probabilities of a target node of any Jackson network was obtained. The formula (that will be given at the end of the next section) led to small values of factor f_V in most cases, including big networks, networks with strong feedback and networks with high dependency. The most problematic cases were networks with high dependency of the target node on the queue length of the other nodes and with the load of the target queue much lower than the load of the other queues.

3 SYSTEM UNDER STUDY

A network with any number of nodes is studied. Customer arrivals and departures are allowed in all the nodes. After being served in node l, customers can go to any node m with probabilities p_{lm} or they can leave the network with probability p_{l0}. The steady-state probability of the number of customers exceeding a level at a target node, $Q_{tg} \ge L$, is estimated. Let us denote K the number of nodes at distance 1, that is, the nodes which are directly connected with the target node ($p_{ltg} \ne 0$), H the number of nodes which are connected with the target node through only one intermediate node (nodes at distance 2), and N the number of nodes at distance 3, for any value of K, H and N. The nodes at distance greater than 2, 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. Customers with independent Poisson arrivals enter each node from the outside with arrival rates $\gamma_{0n}, n = 1, \ldots, N$ to the nodes at distance 3, $\gamma_{1i}, i = 1, \ldots, H$ to the nodes at distance 2, $\gamma_{2j}, j = 1, \ldots, K$ to the nodes at distance 1 and γ_{tg} to the target node. The total arrival rates to each node (arrivals from the outside + arrivals from the other nodes) are denoted by: $\lambda_{0n}, n = 1, \ldots, N$, $\lambda_{1i}, i = 1, \ldots, H$, $\lambda_{2j}, j = 1, \ldots, K$ and λ_{tg}, respectively. The service times at all the nodes are assumed to have an Erlang distribution with α phases (2 or 3 in the examples of Section 4) and with service rates $\mu_{0n}, n = 1, \ldots, N$, $\mu_{1i}, i = 1, \ldots, H$, $\mu_{2j}, j = 1, \ldots, K$ and μ_{tg}, respectively. The buffer space in each queue is assumed to be infinite. Let us observe that $\lambda_{1i} = \gamma_{1i} + \sum_{n=1}^{N} \lambda_{0n} p_{ni} + \sum_{l=1}^{H} \lambda_{1l} p_{li} + \sum_{j=1}^{K} \lambda_{2j} p_{ji} + \lambda_{tg} p_{tgi}, i = 1, \ldots, H$, $\lambda_{2j} = \gamma_{2j} + \sum_{i=1}^{H} \lambda_{1i} p_{ij} + \sum_{l=1}^{K} \lambda_{2l} p_{lj} + \lambda_{tg} p_{tgj},\ j = 1, \ldots, K$, $\lambda_{tg} = \gamma_{tg} + \sum_{j=1}^{K} \lambda_{2j} p_{jtg} + \lambda_{tg} p_{tgtg}$. The loads of the nodes are $\rho_{1i} = \lambda_{1i} / \mu_{1i}, i = 1, \ldots, H$, $\rho_{2j} = \lambda_{2j} / \mu_{2j}, j = 1, \ldots, K$ and $\rho_{tg} = \lambda_{tg} / \mu_{tg}$, respectively.

A formula of the importance function for estimating the probability that $Q_{tg} \ge L$ for any Jackson network was obtained in Villén-Altamirano (2009). With that formula, very low probabilities for different network topologies were accurately estimated within short or moderate computational times. A simplified version of this formula (also given in that paper) that matches with the general one in almost all cases is the following:

$$\Phi = \sum_{i=1}^{H} \alpha_{1i} \frac{\ln\left(\rho_{tg} / \rho_{tgi}^{*}\right)}{\ln \rho_{tg}} Q_{1i} + \sum_{j=1}^{K} \alpha_{2j} \frac{\ln\left(\rho_{tg} / \rho_{tgj}^{\perp}\right)}{\ln \rho_{tg}} Q_{2j} + Q_{tg}, \tag{4}$$

where:

$$\rho_{tgi}^{*} = \rho_{tg} \frac{\gamma_{tg} + \sum_{j=1}^{K} Min\left\{\lambda_{2j} + \left(\mu_{1i} - \lambda_{1i}\right) p_{ij}, \mu_{2j}\right\} p_{jtg} + \lambda_{tg} p_{tgtg}}{\lambda_{tg}}$$

$$\rho_{tgj}^{\perp} = \frac{\gamma_{tg} + \mu_{2j} p_{jtg} + \sum_{l \neq j} \lambda_{2l} p_{ltg} + \lambda_{tg} p_{tgtg}}{\mu_{tg}} = \rho_{tg} \left(1 + \frac{\left(\mu_{2j} - \lambda_{2j}\right) p_{jtg}}{\lambda_{tg}}\right)$$

$$\alpha_{1i} = 1 + \frac{\sum_{l \neq i} \gamma_{1l} \sum_{j=1}^{K} p_{lj} p_{jtg} + \sum_{j=1}^{K} \gamma_{2j} p_{jtg} + \gamma_{tg}}{\mu_{1i} \sum_{j=1}^{K} p_{ij} p_{jtg}} \; ; \; \alpha_{2j} = 1 + \frac{\sum_{i=1}^{H} \gamma_{1i} \sum_{l \neq j} p_{il} p_{ltg} + \sum_{l \neq j} \gamma_{2l} p_{ltg} + \gamma_{tg}}{\mu_{2j} p_{jtg}}.$$

ρ_{tgi}^{*} and $\rho_{tgj}^{\perp}$ are, approximately, the loads of the target queue when a node *i* at distance 2 from the target node or a node *j* at distance 1, respectively, are not empty. It is more difficult to get insight of the meaning of α_{1i} and α_{2j} without following the derivation of Equation (4). Nevertheless, the formulas are easy to apply because all their terms are parameters of the system.

This importance function was derived equating the importance of one extreme state (a system state when only one queue is not empty) with the importance of each of the other extreme states. We will study whether the importance of the extreme states would be affected in a similar manner when the services times are not exponentially distributed and, as a consequence, the importance function derived for Jackson networks would be fit for networks with Erlang service times.

As a particular case of the previous model, we will study a three-queue tandem network with the third node as the target node. For this network, Equation (4) matches with the following equations:

$$\text{If } \rho_3 < \rho_2 < \rho_1 \text{ then } \Phi = \frac{\ln \rho_1}{\ln \rho_3} Q_1 + \frac{\ln \rho_2}{\ln \rho_3} Q_2 + Q_3. \tag{5}$$

$$\text{If } \rho_1 \le \rho_3 < \rho_2 \text{ or if } \rho_3 \le \rho_1 < \rho_2 \text{ then } \Phi = \frac{\ln \rho_2}{\ln \rho_3} \left(Q_1 + Q_2\right) + Q_3. \tag{6}$$

$$\text{If } \rho_1 \le \rho_2 \le \rho_3 \text{ or if } \rho_2 \le \rho_1 \le \rho_3 \text{ then } \Phi = Q_1 + Q_2 + Q_3. \tag{7}$$

$$\text{If } \rho_2 < \rho_3 < \rho_1 \text{ then } \Phi = \frac{\ln \rho_1}{\ln \rho_3} Q_1 + Q_2 + Q_3.$$

Remark: Φ is a function of the current state of the system that could change at each instant *t*. This dependence on *t* could be written in the formulas. For example, Equation (5) could be written as: $\Phi(t) = \frac{\ln \rho_1}{\ln \rho_3} Q_1(t) + \frac{\ln \rho_2}{\ln \rho_3} Q_2(t) + Q_3(t)$.

4 TEST CASES

We conducted several simulation experiments on networks with different topologies and loads. The rare set *A* was defined as $Q_{tg} \ge L$, where Q_{tg} is the number of customer at the target node. The steady state probability of *A* was of the order of 10^{-15} in all the examples. Thresholds T_i were set for every integer value of Φ between 2 and *M*, where *M* varies between *L* and *L*+2 depending on the case being simulated. Therefore, the rare set can be reached in retrials from the last 2, 3 or 4 regions $C_i - C_{i+1}$ (C_M if *i* = *M*). Pilot runs (one or two for each case) were made to estimate the probabilities $P_{i/0}, i = 1, ..., M+1$, and then to set the number of retrials according to Eq. (2), following the guidelines given in Section 2 for rounding to integer values. The interval width of each confidence interval was evaluated using the independent replication method. Each replication (sample) finished after a fix number of arrivals (usually between 100.000 and 500.000). After each sample the half width of the 95% confidence interval divided by the estimate (relative error) was calculated and the simulation finished when the relative error was smaller than 0.1. For each case we made 3 simulations, and we wrote in the tables the results corresponding to the median of the computational times. All the experiments were run on a Pentium(R) D CPU 3.01 GHz.

4.1 Three-Queue Tandem Network

Customers arrive to the first queue of this network according to a Poisson distribution with arrival rate equal to 1, then go to the second queue, then to the third and then they leave the network. The service time at each node follows an Erlang distribution with shape parameter equal to α (2 or 3). Initially, the importance function Φ was chosen according to Eqs. (5), (6), and (7), that is, $\Phi \ = \ aQ_1 + bQ_2 + Q_3$, with values of *a* and *b* depending on the loads ρ_1, ρ_2 and ρ_3. Then, the coefficients *a* and *b* of Φ were multiplied by a correction factor *k*, for *k* = 0.6, 0.7, 0.8 and 0.9. The comparison among the importance functions was made for less rare set probabilities (of the order of 10^{-9}) because the results obtained are also valid for much lower probabilities, and so we could do the comparison with shorter computational effort. We have checked that the best value of *k* for estimating probabilities of the order of 10^{-9} was also the best for estimating probabilities of the order of 10^{-15}. The results, summarized in Table 1, correspond to the values of *a* and *b* for which the computational time was lowest.

Table 1: Results for the three-queue tandem network with Poisson arrivals and Erlang (α, β) service times. Rare set probability: $P\left(Q_3 \geq L\right)$. Relative error = 0.1. $\rho_3 = 1/3$. $\Phi \ = \ aQ_1 + bQ_2 + Q_3$.

ρ_1	ρ_2	α	L	$\hat{P}$	k	a	b	Events (millions)	Time (minutes)	Gain (events)	f_V	Gain (time)	f_0
2/3	1/2	2	18	$2.18\cdot10^{-15}$	0.7	0.26	0.44	13.5	9.6	$1.1\cdot10^{10}$	24.9	$7.6\cdot10^{8}$	14.5
2/3	1/2	3	13	$5.21\cdot10^{-15}$	0.6	0.22	0.38	23.0	21.8	$3.6\cdot10^{9}$	30.9	$2.4\cdot10^{8}$	15.0
1/3	1/2	2	20	$1.92\cdot10^{-15}$	0.8	0.50	0.50	2.9	1.8	$1.2\cdot10^{11}$	2.6	$9.4\cdot10^{9}$	12.8
1/3	1/2	3	16	$9.17\cdot10^{-16}$	0.7	0.44	0.44	4.1	3.9	$8.1\cdot10^{10}$	7.6	$5.6\cdot10^{9}$	14.5
1/5	1/4	2	24	$1.72\cdot10^{-15}$	0.9	0.9	0.9	0.6	0.4	$4.8\cdot10^{11}$	0.7	$3.6\cdot10^{10}$	13.4
1/5	1/4	3	21	$2.99\cdot10^{-15}$	0.8	0.8	0.8	0.8	0.6	$1.3\cdot10^{11}$	1.6	$9.3\cdot10^{9}$	14.0

We can observe that for $\alpha = 2$, the value of *k* is closer to 1 than for $\alpha = 3$. As the coefficient of variation of Pearson of the Erlang distribution is $1\big/\sqrt{\alpha}$, it seems that the more similar is this coefficient to that of the exponential distribution, the closer is the importance function to that given by the formula. It is also observed that lower values of the loads of the two first nodes lead to importance functions closer to those derived for the exponential distribution. Future studies will give us more insight for estimating in advance a good value of *k*.

To evaluate approximately the gain in events or the gain in time with respect to a crude simulation, the data of the crude simulations were estimated by extrapolating the measured values for lower probabilities, see e.g., Villén-Altamirano and Villén-Altamirano (2006). The extrapolation was made taking into account that with crude simulation the relative error for estimating a probability *P* with *n* samples is proportional to $\sqrt{P\left(1-P\right)/n}\big/P$ □ $1\big/\sqrt{nP}$. Thus, the number of samples for achieving a given relative error is inversely proportional to the probability that is to be estimated. Factors f_V is estimated by comparing the measured gain in events with the theoretical one derived from Eq. (1), following the procedure explained in that paper and f_O is estimated as the ratio between gain in events and gain in time.

The values of factor f_O are much greater than those obtained in Villén-Altamirano (2009) for the same network but with exponential service times. The reason is that rescheduling is straightforward only for the exponential distribution due to the memoryless property of this model. Rescheduling service times of any other distribution is more time consuming. We can proceed as follows: a random value of the whole service time of a costumer is obtained. If that value is greater than the service time at the current time, the remaining service time of the customer is obtained as the difference between the two amounts. Otherwise a new random value is obtained and so on. As we must reschedule the service times when any threshold is reached, the computational time with Erlang service times is around four times greater than with exponential times for estimating a probability of the order of 10^{-15}.

The values of factor f_V are similar to those obtained in Villén-Altamirano (2009) for the same network but with exponential service times. The low or moderate values of f_V show that the choice of the importance function is appropriate and that the application is not far from the optimal, at least for the tested cases. We observe that the worst results (greatest computational times and greatest values of factor f_V) were obtained when $\rho_3 < \rho_2 < \rho_1$, but even in these cases the computational times are moderate (lower than 22 minutes). For the other four cases, the computational times needed for estimating probabilities of the order of 10^{-15} were lower than 4 minutes.

We have studied the sensitivity of the results to changes in the importance function, in particular for various correction factors. We have observed that values of f_V for values of a and b up to 10% (20% in the last two cases) lower or greater than those given in Table 1, became less than double those of the best result. The importance functions given by Eqs. 5-7 lead to greater computational times. In the last four cases these times are moderates (lower than 30 minutes), but in the second case the computational time was greater than one day.

4.2 A Network with Seven Nodes

Let us now consider a network with 7 nodes. Customers from the outside arrive at any node of the network with a rate $\gamma_i = 1,\ i = 1,\ldots,7$. After being served in each node, a customer leaves the network with probability 0.2. Otherwise, the customer goes to another node in accordance to the following transition matrix:

	1	2	3	4	5	6	tg
1	0.1	0.1	0.1	0.1	0.2	0.2	0
2	0.1	0.1	0.1	0.1	0.2	0.2	0
3	0.1	0.1	0.1	0.1	0.2	0.2	0
4	0.1	0.1	0.1	0.1	0.2	0.2	0
5	0.1	0.1	0.1	0.1	0	0.1	0.3
6	0.1	0.1	0.1	0.1	0.1	0	0.3
tg	0.1	0.1	0.1	0.1	0.1	0.1	0.2

We can observe in the matrix that nodes 5 and 6 are at distance 1 from the target node, and that nodes 1 to 4 are at distance 2 from the target node. We can also observe that the network is a full mesh network but only nodes 5 and 6 are directly connected to the target. The overall arrival rate to each node is $\lambda_i = 4.5$ for nodes at distance 2, $\lambda_j = 5.73$ for nodes at distance 1 and $\lambda_{tg} = 5.55$. The rare set is $Q_{tg} \geq L$, where Q_{tg} is the target queue. The load of the target node is 0.3262. Services rates were chosen such that nodes 5 and 6 have the same load, given by ρ_2 in Table 2, and also nodes 1 to 4 (ρ_1 in Table 2).

The importance function Φ was chosen according to Eq. (4) with the coefficients given in Section 3, that is, without applying any correction factor. The gain and the factors f_V and f_O were estimated as in Section 4.1. The results, which are summarized in Table 2, show that very accurate results were obtained in short computational time, less than 2 minutes in all the cases.

Table 2: Results for a Jackson network with 7 nodes, Poisson arrivals and Erlang (α, β) service times. Rare set probability: $P\left(Q_{tg} \geq L\right)$. Relative error = 0.1. $\rho_{tg} = 0.3262$. $\Phi = a\sum_{i=1}^{4} Q_i + b\sum_{j=5}^{6} Q_j + Q_{tg}$.

ρ_1	ρ_2	α	L	$\hat{P}$	a	b	Events (millions)	Time (seconds)	Gain (events)	f_V	Gain (time)	f_O
0.50	0.41	2	26	$2.69\cdot10^{-15}$	0.23	0.45	2.2	77	$1.9\cdot10^{11}$	1.4	$2.9\cdot10^{10}$	6.6
0.50	0.41	3	25	$1.85\cdot10^{-15}$	0.23	0.45	2.7	93	$2.2\cdot10^{11}$	1.7	$3.2\cdot10^{10}$	6.8
0.32	0.41	2	26	$3.80\cdot10^{-15}$	0.36	0.45	1.3	40	$2.1\cdot10^{11}$	0.9	$3.6\cdot10^{10}$	5.8
0.32	0.41	3	25	$2.76\cdot10^{-15}$	0.36	0.45	2.1	66	$1.7\cdot10^{11}$	1.5	$2.9\cdot10^{10}$	5.8
0.28	0.30	2	27	$1.95\cdot10^{-15}$	0.40	0.61	1.3	40	$4.4\cdot10^{11}$	0.8	$7.2\cdot10^{10}$	6.1
0.28	0.30	3	26	$1.73\cdot10^{-15}$	0.40	0.61	1.8	59	$3.0\cdot10^{11}$	1.3	$4.8\cdot10^{10}$	6.2

We can also observe that the worst results (greater computational times) were obtained when $\rho_{tg} < \rho_2 < \rho_1$, though the computational times needed for estimating probabilities of the same order of magnitude is much lower than in the previous network. The results are better than in this network because many customers of the other queues never go to the target queue. Thus, the dependence of the target queue on the queue length of the other queues is weaker and the efficiency of RESTART is greater. Unlike with importance sampling, the efficiency of RESTART may improve with the complexity of the systems and it does not seem affected by the feedback.

The values of factor f_O are also greater than those obtained with exponential service times, but the differences are lower than in the previous network because in complex networks the proportion of computational time that the simulation program is making rescheduling is lower than in simple networks.

In the three cases of $\alpha = 2$ the best results were obtained with the importance function given by Eq. (4), while for $\alpha = 3$ the best results were obtained with coefficients of Q_{1i} and Q_{2j} around 10% lower than those given by Eq. (4), that is, with a correction factor of $k = 0.9$. Nevertheless, very good results were also obtained without any correction factor, as can be seen in the table. The very low values of f_V achieved show that the application is very close to the optimal one, at least for the tested cases. The sensitivity of the choice of these coefficients is smaller than in the previous network, because acceptable results (values of f_V less than double those of the best results) are obtained for values of a and b up to 20% lower or greater than the optimal ones.

The three cases were also simulated for $\alpha = 10$ with the importance function given by Eq. (4). The parameter β of the Erlang distribution was also changed for achieving the loads of Table 2. The computational times for estimating probabilities of the order of 10^{-15} were very similar to those obtained with $\alpha = 3$. It seems to indicate that this importance function leads to very good results in this network for Erlang service times with any number of phases α.

5 CONCLUSIONS

The choice of the importance function is the most critical feature when the method RESTART is applied to multi-dimensional systems. This paper has focused on finding formulas of effective importance functions for two types of networks: a three-queue tandem network and a more complex network with seven nodes, in both cases with different loads and with Poisson arrivals and Erlang service times. In the second network the formulas of the importance functions were the same as those derived in Villén-Altamirano (2009) for exponential service times, while for the tandem network the coefficients of the importance functions were multiplied by a correction factor equal to 0.6, 0.7, 0.8 or 0.9. This correction factor is closer to 1 for $\alpha = 2$ than for $\alpha = 3$. It could be due to the coefficient of variation of Pearson that is more similar to that of the exponential distribution for $\alpha = 2$. Although it has been easy to obtain heuristically the correction factor, future studies will give us more insight for estimating in advance a good value of this factor.

Overflow probabilities lower than those needed in practical problems (around 10^{-15}) have been accurately estimated within short computational work. The worst results are obtained when the dependence of the target queue on the length of the other queues is very high (as it occurs in a tandem network) and the load of the target queue is much lower than the other ones. As a consequence, the efficiency of RESTART often improves with the complexity of the system because the dependence of the target queue on the other queues is weaker. The efficiency does not seem affected by the feedback.

It would be interesting to see whether the importance function derived for Jackson networks would be fit for other networks or at least if a suitable importance function can be obtained by multiplying the coefficients of the importance function (except the target coefficient) by the same correction factors. The methodology followed in this paper could be used in future researches to determine the types of non-Markovian multi-dimensional networks for which those importance functions are valid.

ACKNOWLEDGEMENTS

I should like to thank Mr. Iván García-Jiménez for his assistance in writing the simulation program and Mr. Manuel Villén-Altamirano for his helpful comments.

REFERENCES

Dean, T., and P. Dupuis. The design and analysis of a generalized DPR/RESTART algorithm for rare event simulation. Submitted to *Annals of Operation Research.*

Dupuis, P., and H. Wang. Importance sampling for Jackson networks. Submitted to *QUESTA*.

Garvels, M.J.J., J.K.C.W. Ommeren, and D.P. Kroese. 2002. On the importance function in splitting simulation. *European Transaction on Telecommunications* 13(4): 363-371.

Glasserman, P., P. Heidelberger, P. Shahabuddin, and T. Zajic. 1999. Multilevel splitting for estimating rare event probabilities. *Operation Research* 47(4): 585-600.

Haraszti, Z., and J.K. Townsend. 1999. The theory of direct probability redistribution and its application to rare event simulation. *ACM Transaction on Modelling and Computer Simulation* 9(2): 105-140.

Kahn H, and T.E. Harris. 1951. Estimation of particle transmission by random sampling. *National Bureau of Standards Applied Mathematics Series* 12: 27-30.

Rubino, G., and B. Tuffin (editors). 2009. R*are even simulation using Monte Carlo methods.* Chichester: Wiley.

Villén-Altamirano J. 2009. Importance functions for RESTART simulation of general Jackson networks. To appear in *European Journal of Operation Research.*

Villén-Altamirano M., A. Martínez-Marrón,, J. Gamo, and F. Fernández-Cuesta 1994. Enhancement of the accelerated simulation method RESTART by considering multiple thresholds. *Proceeding of the 14th International Teletraffic Congress*. In *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks*, 787-810. Elsevier Science Publisher, Amsterdam.

Villén-Altamirano M., and J. Villén-Altamirano. 1991. RESTART: a method for accelerating rare event simulations. *Proceeding of the 13th International Teletraffic Congress.* In *Queueing, Performance and Control in ATM*, 71-76. Elsevier Science Publisher, Amsterdam.

Villén-Altamirano M., and J. Villén-Altamirano. 2002. Analysis of RESTART simulation: theoretical basis and sensitivity study. *European Transaction on Telecommunications* 13(4): 373-386.

Villén-Altamirano M., and J. Villén-Altamirano. 2006. On the efficiency of RESTART for multidimensional state systems. *ACM Transaction on Modelling and Computer Simulation* 16(3): 251-279.

AUTHOR BIOGRAPHY

JOSÉ VILLÉN-ALTAMIRANO is Professor of the Department of Applied Mathematics at the Polytechnic University of Madrid. He received M.S. degree in Mathematics from Complutense University of Madrid in 1977 and he received Ph. D. degree in Computer Science from Polytechnic University of Madrid in 1988. His research interests are focused on reliability, queueing theory and rare event simulation. His email is jvillen@eui.upm.es.