ITC19/ Performance Challenges for Efficient Next Generation Networks 1707-1716
LIANG X.J. and XIN Z.H.(Editors)
V.B. IVERSEN and KUO G.S.(Editors)
Beijing University of Posts and Telecommunications Press

Efficient estimation of blocking probabilities in non-stationary loss networks*

Richard J. Boucherie and Pasi Lassila

[1] University of Twente
Department of Applied Mathematics, P.O.Box 217, 7500 AE Enschede, The Netherlands
R.J.Boucherie@utwente.nl
[2] Helsinki University of Technology
Networking Laboratory, P.O.Box 3000, FIN 02015-HUT, Finland
Pasi.Lassila@hut.fi

Abstract. We consider estimation of blocking probabilities in a nonstationary loss network. By invoking the so called MOL (Modified Offered Load) approximation, the problem is transformed into one requiring the solution of blocking probabilities in stationary loss networks with time varying loads. To estimate these blocking probabilities, Monte Carlo simulation is used and to increase the efficiency of the simulation, we develop a likelihood ratio method that enables samples drawn at one time point to be used at later time points. This reduces the need to draw new independent samples at every time point, thus giving substantial savings in the computational effort. The accuracy of the method is analyzed by using Taylor series approximations of the variance indicating the direct dependence of the accuracy on the rate of change of the actual load. Finally, two practical applications are provided to demonstrate the efficiency of the method.

Keywords: nonstationary loss network, blocking probabilities, Modified Offered Load, Monte Carlo methods, importance sampling

1 INTRODUCTION

Loss networks (see, e.g., [1]) are classical teletraffic models that have been used to evaluate the performance of traditional circuit switched networks. Such circuit switched systems include the existing fixed telephone networks, as well as cellular networks. The traditional loss network model uses the assumption that the arrival process of calls into the network can be described by a time homogeneous Poisson process. However, this assumption is not always justified. Consider, e.g., the need to assess the impact of televoting on the fixed network or traffic jams moving along a highway in cellular networks. To this end, the loss network model can be extended such that the calls of a given traffic class are assumed to be generated by a Poisson process with a time varying rate.

* This research is partly supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs The Netherlands.

In loss networks, one of the basic tasks is to calculate the blocking probability for each traffic class in the system. The steady state distribution of the system in the stationary (time homogeneous) case has the well known product form, from which it is easy to obtain analytic expressions also for the blocking probabilities. In the nonstationary loss network the product form does not hold anymore and hence, the time dependent blocking probabilities do not have explicit analytic expressions, either. However, by using the Modified Offered Load (MOL) approximation, as given in [2], accurate approximations of the blocking probabilities can be obtained. The MOL approximation corresponds to a problem of determining blocking probabilities for a stationary loss network with an offered load depending on time. However, it is well known that exact evaluation of blocking probabilities for stationary loss networks representing realistic size networks is computationally very difficult due to the huge size of the state space. As an alternative, Monte Carlo (MC) simulation can be used to obtain estimates of the blocking probabilities.

This paper addresses the problem of efficiently estimating the MOL approximation by using MC simulation augmented with ideas from the importance sampling (IS) variance reduction method. The objective is to minimize the number of time points for which sampling is required. Ideally, with our *likelihood ratio method* samples are generated only once and these are used for all time points, whereas the direct method of computing the MOL approximation using MC simulation entails generating the samples independently at every time point. To be able to use samples generated at some reference time point at other time points requires the samples to be weighted appropriately with the likelihood ratio. This is similar to IS in stationary loss networks, as studied in [3], [4], [5], and [1], where IS is used to provide better sampling of the most important parts of the state space. In our case, the likelihood ratio is used to scale the result to a different point in time. The method does not incur much extra effort in the simulation and the sample statistics can be used to determine if accuracy is good enough at each time point. When this is not the case, samples are redrawn for such time points. We also compute approximations to the variance of our estimator to obtain insight into the dependence of the variance on the load. Finally, we consider two application scenarios where the likelihood ratio method is used 1) to estimate the blocking probability over a given finite time horizon assuming that the load is also known, and 2) in an on-line algorithm that only utilizes local information of the load and the samples are used until the accuracy criterion fails.

The paper is organized as follows. Section 2 introduces the non-stationary loss network and the MOL approximation. In section 3, we present the likelihood ratio method and our analytical results. Applications are given in Section 4, and the conclusions in Section 5.

2 THE NONSTATIONARY LOSS NETWORK

Consider a network having J links, indexed $j = 1, \ldots, J$, with link j having a capacity of C_j resource units. The network supports K classes of calls. A class k call, $k = 1, \ldots, K$, has a bandwidth requirement of $b_{j,k}$ on link j, $b_{j,k} = 0$ when class k does not use link j. The vector $\mathbf{b}_j$ denotes the required bandwidths of all classes on link j. Calls of class k arrive according to an inhomogeneous Poisson process with time dependent arrival rate $\lambda_k(t)$, and have an exponentially distributed holding time with mean $1/\mu_k$. New calls are always accepted if there is enough capacity and blocked calls are cleared.

dependent distribution of the infinite capacity system. Analytical expressions and bounds on the error of the MOL approximation can be found in [2] for a network with unit size calls and a single link, and in [7] for general call sizes and multiple links.

The primary performance measure we are interested in is the instantaneous blocking probability of a class k call at time t, $B_k(t)$. It is the probability that an arriving class k call is blocked at time t and is given by

$$B_k(t) = \mathrm{P}\{\mathbf{X}(t) \in \mathcal{B}_k\}.$$

No explicit analytical expressions exist for computing $B_k(t)$. However, by invoking the MOL approximation, $B_k(t)$ can be expressed in the form of a ratio of two state sums

$$B_k(t) \approx \mathrm{P}\{\mathbf{X}^\infty(t) \in \mathcal{B}_k \,|\, \mathbf{X}^\infty(t) \in \mathcal{S}\} = \frac{\mathrm{P}\{\mathbf{X}^\infty(t) \in \mathcal{B}_k\}}{\mathrm{P}\{\mathbf{X}^\infty(t) \in \mathcal{S}\}}. \tag{3}$$

Computing the MOL approximation at time t consists of computing $\rho_k(t), \forall k$, and then computing the loss probabilities from the time homogeneous loss network with load $\rho_k(t)$. Note that an estimate of $\rho_k(t), \forall k$, as measured from a live network, is sufficient, as well.

All methods available for loss networks (see, e.g., [1] for an overview) can be used to evaluate the blocking probabilities for a given fixed t. Exact computation of these blocking probabilities can be done efficiently only for networks with special topologies and, in practice, approximations are required. Here we investigate the use of so called static Monte Carlo methods to obtain estimates of the blocking probabilities.

3 THE LIKELIHOOD RATIO METHOD

In the following we assume that $B_k(t)$ is to be estimated for a given traffic class k. Thus, the index k is assumed implicit and is omitted, i.e., we denote $B_k(t) \equiv B(t)$, etc. Additionally, note that the simulation requires estimation of two state sums

$$\beta(t) = \mathrm{P}\{\mathbf{X}^\infty(t) \in \mathcal{B}\}, \quad \text{and} \quad \gamma(t) = \mathrm{P}\{\mathbf{X}^\infty(t) \in \mathcal{S}\}.$$

Of these, it is the estimation of $\beta(t)$ that is often inefficient due to the rarity of the blocking states. The probability $\gamma(t) \approx 1$ and is hence easy to estimate, as discussed, e.g., in [8]. Thus, in the sequel, we focus on methods for estimating $\beta(t)$.

Consider estimating the MOL approximation of $\beta(t)$ for $t \in [0, T]$. In the straight forward method, one essentially performs a separate simulation at every time point $t_i \in [0, T]$. For each t_i the standard MC method is used and $\beta(t_i)$ is estimated by $\hat{\beta}_N(t_i) = (1/N) \sum_{n=1}^{N} 1(\mathbf{X}_n^\infty(t_i) \in \mathcal{B})$, where $\mathbf{X}_n^\infty(t_i)$ denotes i.i.d. samples drawn from (1) and N is the total number of samples. This assumes that samples are independently drawn for each time point. The idea in our *likelihood ratio method*, introduced below, is to minimize the number of samples drawn in the simulations when estimating $\beta(t), t \in [0, T]$, by allowing some controlled reduction of accuracy of the estimates at the different time points. Ideally, we want to draw the samples only once, and with those samples estimate $\beta(t_i)$ for all t_i.

To achieve this, ideas from the importance sampling (IS) variance reduction method are applied. IS is based on the following well known property. Consider a discrete random variable X with distribution $p(x)$, and the probability of a set $\mathcal{A}$, denoted by α, which can be expressed as $\alpha = \mathrm{E}[1(X \in \mathcal{A})]$. Introducing another distribution $p^*(x)$ for X (satisfying $p^*(x) > 0, \forall x \in \mathcal{A}$) allows us to write $\alpha = \mathrm{E}_{p^*}[1(X \in \mathcal{A})(p(X)/p^*(X))]$, where $\mathrm{E}_{p^*}[\cdot]$ denotes

where $\mathrm{err}(t^*,t)$ denotes the error terms of the Taylor series expansion of $\beta(t)$ around t^*. Using the Taylor series expansion of $L(\mathbf{x},t^*,t)$ we can alternatively express $\beta(t)$ as

$$\begin{aligned}\beta(t) &= \mathrm{E}_{\rho^*}[1(\mathbf{X}^*\in\mathcal{B})\,L(\mathbf{X}^*,t^*,t)]\\ &= \mathrm{E}_{\rho^*}\Big[1(\mathbf{X}^*\in\mathcal{B})\,\Big(L(\mathbf{X}^*,t^*,t^*)+\tfrac{dL(\mathbf{X}^*,t^*,t^*)}{dt}(t-t^*)+\mathrm{Err}(\mathbf{X}^*,t^*,t)\Big)\Big],\\ &\approx \beta(t^*)+\mathrm{E}_{\rho^*}\Big[1(\mathbf{X}^*\in\mathcal{B})\,\tfrac{dL(\mathbf{X}^*,t^*,t^*)}{dt}\Big]\cdot(t-t^*),\end{aligned} \tag{7}$$

where $dL(\mathbf{x},t^*,t^*)/dt$ is short hand for $dL(\mathbf{x},t^*,t)/dt$ evaluated at time t^*, and

$$\frac{dL(\mathbf{x},t^*,t)}{dt}=\left(\sum_{k=1}^{K}\left(\frac{x_k}{\rho_k(t)}-1\right)\cdot\frac{d\rho_k(t)}{dt}\right)\cdot L(\mathbf{x},t^*,t).$$

The error term in the Taylor series, $\mathrm{Err}(\mathbf{x},t^*,t)$, equals

$$\mathrm{Err}(\mathbf{x},t^*,t)=\frac{d^2L(\mathbf{x},t^*,\hat{t})}{dt^2}\,\frac{(\hat{t}-t^*)^2}{2}, \tag{8}$$

for some $\hat{t}\in[t^*,t]$. To gain insight into the error, in Remark 1 we relate the expected value of this error to the higher order derivatives of $\beta(t)$.

For the variance we need the 2nd moment of $1(\mathbf{X}^*\in\mathcal{B})\,L(\mathbf{X}^*,t^*,t)$,

$$\begin{aligned}\mathrm{E}_{\rho^*}[1(\mathbf{X}^*\in\mathcal{B})\,(L(\mathbf{X}^*,t^*,t))^2] \approx \beta(t^*)&+2\mathrm{E}_{\rho^*}\Big[1(\mathbf{X}^*\in\mathcal{B})\,\tfrac{dL(\mathbf{X}^*,t^*,t^*)}{dt}\Big](t-t^*)\\ &+\mathrm{E}_{\rho^*}\left[1(\mathbf{X}^*\in\mathcal{B})\,\left(\tfrac{dL(\mathbf{X}^*,t^*,t^*)}{dt}\right)^2\right](t-t^*)^2.\end{aligned} \tag{9}$$

Collecting the terms from (7) and (9), we obtain

$$\begin{aligned}&\mathrm{V}_{\rho^*}[1(\mathbf{X}^*\in\mathcal{B})\,L(\mathbf{X}^*,t^*,t)] \approx \beta(t^*)(1-\beta(t^*))\\ &+2(1-\beta(t^*))\mathrm{E}_{\rho^*}\Big[1(\mathbf{X}^*\in\mathcal{B})\,\tfrac{dL(\mathbf{X}^*,t^*,t^*)}{dt}\Big](t-t^*)+\mathrm{V}_{\rho^*}\Big[1(\mathbf{X}^*\in\mathcal{B})\,\tfrac{dL(\mathbf{X}^*,t^*,t^*)}{dt}\Big](t-t^*)^2.\end{aligned} \tag{10}$$

Thus, we have obtained a characterization of the variance of (4) as a function of terms depending only on random variables at the reference time t^* and the distance at time t from the reference time t^*. The dependencies are such that the variance depends quadratically in time on the variance of $1(\mathbf{X}^*\in\mathcal{B})\,\frac{dL(\mathbf{X}^*,t^*,t)}{dt}$ and linearly in time on the expectation of $1(\mathbf{X}^*\in\mathcal{B})\,\frac{dL(\mathbf{X}^*,t^*,t)}{dt}$. Furthermore, note that (10) is simply a 2nd order equation in t.

Remark 1: Here we comment on the error of the Taylor series approximation. First, it can be observed that the terms in the Taylor series expansions (6) and (7) of $\beta(t)$ satisfy

$$\mathrm{E}_{\rho^*}\left[1(\mathbf{X}^*\in\mathcal{B})\,\frac{d^kL(\mathbf{X}^*,t^*,t^*)}{dt^k}\right]=\sum_{\mathbf{x}\in\mathcal{B}}\frac{1}{\pi(\mathbf{x},t^*)}\frac{d^k\pi(\mathbf{x},t^*)}{dt^k}\pi(\mathbf{x},t^*)=\frac{d^k\beta(t^*)}{dt^k},$$

i.e., the terms of the Taylor series (7) with respect to $L(\mathbf{x},t^*,t)$ coincide term by term with the higher order derivatives of $\beta(t)$ in the Taylor series (6) of β. This also follows from the uniqueness of the Taylor series. Thus, the expected value of the error (8) equals

$$\mathrm{E}_{\rho^*}[1(\mathbf{X}^*\in\mathcal{B})\,\mathrm{Err}(\mathbf{X}^*,t^*,t)]=\mathrm{err}(t^*,t)=\frac{d^2\beta(\hat{t})}{dt^2}\,\frac{(\hat{t}-t^*)^2}{2},$$

for some $\hat{t}\in[t^*,t]$. Further, it can be seen that $d^2\beta(t)/dt^2\sim d^2\rho_k(t)/dt^2,\forall k$ (see [7] and [2]). In conclusion, we note that the error in $\beta(t)$ by using the Taylor series approximation

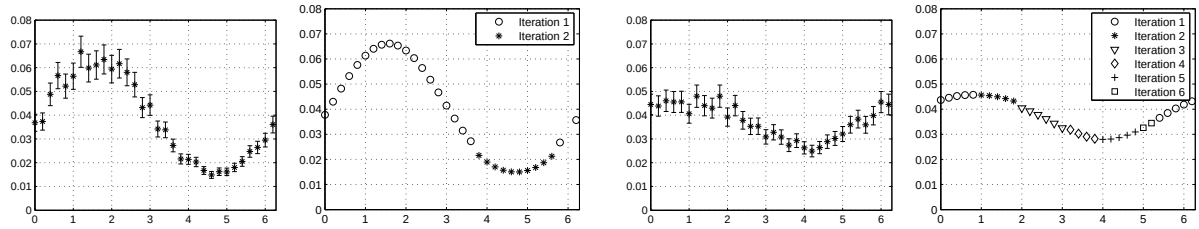

Fig. 1. Results for the uniform load scenario with standard MC (1st fig) and likelihood ratio method (2nd fig), and for the nonuniform load scenario with standard MC (3rd fig) and likelihood ratio method (4th fig).

same way for both the standard MC and the likelihood ratio method. We have used the discretized inverse transformation method, where the necessary Poisson distributions are simply computed into arrays and sample generation corresponds to a table lookup using a linear search.

4.1 MOL approximation for fixed time points

The problem considered here is that given $\rho_k(t)$, for all k, and a set of time points t_m, with $m = 1, \ldots, M$, how should one compute the MOL approximation of $\beta(t_m), \forall m$. This corresponds to computing a discretized approximation to the continuous $\beta(t)$.

To apply the likelihood ratio method, we fix the initial reference load (or time) as $\boldsymbol{\rho}^* = \{\rho_1(t_1), \ldots, \rho_K(t_1)\}$. Then, samples are generated from the distribution (1) with load $\boldsymbol{\rho}^*$ until the relative error (ratio of standard deviation to mean) is less than $\varepsilon_2 \leq \varepsilon_1$, where ε_1 is the target relative error to be met by all time points. Choosing $\varepsilon_2 < \varepsilon_1$ can be used to increase the likelihood that time points having loads that are similar to the reference load will meet the target accuracy ε_1. The same samples are used in estimator (4) to obtain estimates of $B_k(t), \forall k$, and the relative errors of the estimates at all points $t_m, m = 1, \ldots, M$. For those time points not meeting the target accuracy ε_1, new samples are drawn using the load corresponding to the first time point not fulfilling the accuracy criterion. Resampling is performed until all time points meet the accuracy criterion.

We computed $B_1(t)$, the blocking probability of class 1, for 32 evenly spaced time points in the interval $[0, 2\pi]$ for both load scenarios. Two accuracy criterions were used, $(\varepsilon_1 = 0.05, \varepsilon_2 = 0.045)$ and $(\varepsilon_1 = 0.02, \varepsilon_2 = 0.018)$. The results using the likelihood ratio method are compared against results given by direct MC simulation (corresponding to an independent simulation at each time point until an accuracy of ε_1 is reached). Figure 1 shows the results for both methods and for both load scenarios. In more detail, 1st figure from left corresponds to the uniform load scenario and shows our estimate of $B_1(t)$ and the 95% confidence intervals using the standard MC method with $\varepsilon_1 = 0.05$ accuracy criterion. The 2nd figure from left corresponds to the same uniform load scenario and shows the estimate for $B_1(t)$ when using the likelihood ratio method for accuracy criterions $(\varepsilon_1 = 0.05, \varepsilon_2 = 0.045)$. The iteration round at which each time point satisfied the target accuracy ε_1 is also shown in the figure via a change of the marker. The 95% confidence intervals are only slightly shorter in magnitude than in the standard MC case, and have been left out to keep the figure clear. Finally, the 3rd and 4th figures show the same results as the 1st and 2nd figures, but for the nonuniform load scenario. Finally, the total number of generated samples for standard MC and the likelihood ratio methods are shown in Table 1, where also the total

effectively scale the blocking probabilities of one time point into the blocking probabilities of also other time points. In other words, using the likelihood ratio method enables us to reuse the samples generated at one time point at other time points, as well, without the need to draw the samples again (and again) at every considered time point, thus saving potentially a lot of computational effort. To gain insight into what factors affect the performance of the method, the variance of the likelihood ratio estimator has been analyzed by using Taylor series techniques, where it is shown that the variability of the offered load is the key component for prediction of the confidence intervals for the blocking probabilities. Additionally, several variants of the method that can be applied in different practical circumstances have been given. Numerical examples comparing the performance of the likelihood ratio method and the standard MC clearly show the efficiency of the likelihood ratio method.

Topics for future research include improving the efficiency further by deriving an importance sampling distribution that uses more samples from the set of blocking states such that variance is reduced in many time points. This requires analysis of the sensitivity of the most likely blocking state(s) with respect to the changing load.

References

1. Ross, K.W.: Multiservice Loss Models for Broadband Telecommunication Networks. Springer, London (1995)
2. Massey, W.A., Whitt, W.: An analysis of the Modified Offered Load approximation for nonstationary Erlang loss model. Annals of Applied Probability **4** (1994) 1145–1160
3. Lassila, P., Virtamo, J.: Efficient importance sampling for Monte Carlo simulation of loss systems. In: Proceedings of 16th International Teletraffic Congress (ITC-16), Edinburgh, UK (1999) 787–796
4. Lassila, P., Virtamo, J.: Nearly optimal importance sampling for Monte Carlo simulation of loss systems. ACM Transaction on Modeling and Computer Simulation **10** (2000) 326–347
5. Mandjes, M.: Fast simulation of blocking probabilities in loss networks. European Journal of Operational Research **101** (1997) 393–405
6. Massey, W.A., Whitt, W.: Networks of infinite server queues with nonstationary Poisson input. Queuing Systems **13** (1993) 183–250
7. Abdalla, N., Boucherie, R.J.: Blocking probabilities in mobile communications networks with time varying rates and redialing subscribers. Annals of Operations Research **112** (2002) 15–34
8. Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V.F., Glynn, P.W.: Nearly optimal importance sampling for Monte Carlo simulation of loss systems. ACM Transaction on Modeling and Computer Simulation **10** (2000) 326–347
9. Sadowsky, J.S.: On the optimality and stability of exponential twisting in Monte Carlo estimation. IEEE Transactions on Information Theory **39** (1993) 119–128