Estimating End-To-End Delay in ATM Virtual Paths

Nelson Fonseca and John A. Silvester

CENG Technical Report 93-36

Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, California 90089-2562
(213)740-4579

September 1993

ESTIMATING END-TO-END DELAY IN ATM VIRTUAL PATHS

Nelson L.S. Fonseca and John A. Silvester
Department of Electrical Engineering - Systems
University of Southern California
Los Angeles, CA 90089-2562
Phone: (213) 740-4579 Fax: (213) 740-9803

Abstract

In ATM networks supporting B-ISDN, the traffic is highly correlated and neglecting this correlation may lead to severe underestimation of end-to-end delay and loss. The ability to accurately estimate end-to-end performance is of paramout importance for the traffic control. The purpose of this paper is to introduce a procedure for estimating end-to-end delays in ATM virtual paths. We model the flows in ATM network as Discrete Batch Markovian Arrival Processes (D-BMAP) and present a framework for the analysis of queueing networks with markov modulated type of flow. We also show a procedure for modeling the output process of a D-BMAP/D/1/k queue.

I) INTRODUCTION

The future Broadband Integrated Services Digital Network will carry video, voice and data applications with different Quality of Service requirements. The cell arrival stream from integrated traffic is highly correlated and neglecting its correlations may lead to severe underestimation of the delay and loss probability [1]-[3]. Although, many researchers have evaluated the impact of correlated traffic on the performance of an isolated multiplexer [4]-[9], little attention has been given to the estimation of end-to-end performance. In this paper, we introduce a procedure for computing end-to-end delay in ATM networks. We specify a framework for queueing networks with Markov modulated flow and show a procedure for modelling the output process of an ATM switching An ATM multiplexer.

In delay sensitive applications, if a cell arrives at the destination after a certain treshold, it becomes useless for signal reconstruction. While the propagation and transmission delay account for the fixed part of the end-to-end delay, the queueing delay is mainly responsible for the variable component of end-to-end delay (jitter). A precise estimation of end-to-end loss metrics such as loss probability and length of loss gap [10] depends on the accurate estimation of these metrics at each individual node. Finding an appropriate queueing network representation for B-ISDN networks is of paramount importance for the computation of end-to-end values.

We assume that the input traffic to the queueing network is modelled as a Discrete Time Batch Markovian Arrival Process (D-BMAP) [11]. The D-BMAP process has been successfully used for modelling the integration of voice, video and data sources [15]-[16]. An ATM multiplexer is viewed as a finite buffer queue with FCFS services (D-BMAP/D/1/k) . We model the output process of a multiplexer as a two-state Markov Modulated Bernoulli Process (MMBP). Insofar as the MMBP is a sub-case of the more general D-BMAP, we are able to maintain an uniform representation of the flows in a queueing network. We use the parametric decomposition approximation to decompose the network of queues in isolated queues [17]. Our approach can be seen as a generalization of the Queueing network Analyzer [18] approach for queueing networks with

Markov modulated flow

This paper is organized as follows. In section II, we briefly describe the traffic model. Section III details the framework for queueing networks with Markov modulated flow. Section IV shows a matching procedure for modeling the output process of a D-BMAP/D/1/k. In section V, we introduce a procedure for computing end-to-end delay in virtual paths. Section VI validates the computational procedure through numerical examples, and finally some conclusions are drawn in section VII.

II) THE TRAFFIC MODEL

In this section we briefly describe the stochastic processes which are used to represent the traffic flow in our queueing network model. In the Discrete Time Batch Markovian Arrival Process [11], a batch may arrive at every discrete time. The batch size probability mass function depends on the state of an underlying discrete time Markov chain. A D-BMAP is completely specified by the matrices D_n whose elements $(d_{ij})_n$ give the probability that a transition from state i to state j occurs and a batch of size n arrives. Clearly, we have that:

$$\sum_{n=0}^{\infty} \sum_{j=1}^{m} (d_{ij})_n = 1$$

where m is the number of states of the underlying Markov chain.

The matrices D_n are related to the transition probability matrix D of the underlying Markov chain by:

$$D = \sum_{n=0}^{\infty} D_n$$

The Discrete Time Batch Markovian Arrival Process is the discrete time version of the Batch Markovian Arrival Process [12]-[14]. The D-BMAP is a non-renewal process. Such a process is very attractive for modelling integrated traffic due to the flexibility that we have to set the parameters $((d_{ij})_n)$ to generate a specific correlation pattern [15]-[20].

Wang and Silvester [15] developed an accurate procedure for matching the statistics of integrated (video, voice and data) traffic with the statistics of a two state D-BMAP. Blondia [16] showed how to map the short and long-term correlations of a video source into a D-BMAP. In [11], Blondia provides a comprehensive description of the D-BMAP process and specifically of its counting process. Hashida et al. [19] derived the statistics of both the counting and interarrival processes for the specific case where the underlying Markov Chain has only two states. In [11], it was demonstrated that the mean arrival rate, the variance of number of arrivals and the covariance at lag k are given by:

$$\lambda = \pi\left(\sum_{k=1}^{\infty} kD_k\right)\bar{e}$$

$$var = \pi\left(\sum_{k=1}^{\infty} k^2 D_k\right)\bar{e} - \lambda^2$$

$$cov(x_1, x_k) = \pi\left(\sum_{n=1}^{\infty} nD_n\right)D^{k-2}\left(\sum_{n=1}^{\infty} nD_n\right)\bar{e} - \lambda^2$$

where $\bar{e}$ is the unit column vector and π is the steady state probability of the underlying Markov chain, i.e;

$$\pi D = \pi \qquad \pi\bar{e} = 1$$

The Markov Modulated Bernoulli Process (MMBP) is a subcase of D-BMAP of interest in this paper. In a Markov Modulated Bernoulli Process, at most a single cell may arrive at each time epoch. We focus our attention on the two-state MMBP which is totally specified by $(p_1, p_2, \alpha_1, \alpha_2)$ where p_i $(i=1,2)$ is the probability of having an arrival when the underlying Markov chain is at state i, and α_i (i=1,2) is the probability of remaining in state i at each time epoch. The matrices D_n of a two-state MMBP are given by:

$$D_0 = \begin{bmatrix} (1-p_1)\alpha_1 & (1-p_1)(1-\alpha_1) \\ (1-p_2)(1-\alpha_2) & (1-p_2)\alpha_2 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} p_1\alpha_1 & p_1(1-\alpha_1) \\ p_2(1-\alpha_2) & p_2\alpha_2 \end{bmatrix}$$

The Markov Modulated Bernoulli Process has also been widely applied to the modelling of integrated traffic [21]-[23]. Le Boudec used the MMBP as input traffic to evaluate the cell loss rate in a multiplexer with buffer priority [21]. Guillermin et al. [22] studied the burstiness concept in B-ISDN networks via an MMBP model. They solved the MMBP/D/1/K and proved a relationship between the stationary queue length distribution and the arrival time distribution in a MMBP/D/1/K queue.

II) QUEUEING NETWORKS WITH MARKOV MODULATED FLOWS

The global flow of a communication network is normally described by queueing network models. In the so-called product form networks, the flow is (or is assumed to be) Poisson and the probability mass function of the distribution of customers among the network nodes can be computed as the product of the probability mass function of the number of customers at each node [24]. Although the product form assumptions allowed the development of computationally efficient algorithms, they are too restrictive to be used in many practical situations [25]. Alternatively, a more realistic approach is to represent the flow in communication networks as a renewal process specified by the mean and variance of the number of arrivals [17]-[26]. The parametric decomposition approximation is used in this approach [17]. The parametric decomposition evaluates each queue in the network as if they were stochastically independent. The queues are analyzed in isolation only after the input flow parameters (mean and variance) are computed via renewal approximations. The parametric decomposition can be seen as a generalization of the product form concept in which the dependencies among the queues are captured by the estimation of the flow parameters. Whitt used the parametric decomposition together with renewal approximations of point processes to develop the Queueing Network Analyzer [18].

In the future B-ISDN network, the traffic will be highly correlated; a renewal representation of the network flow can no longer be accepted [27]. Recently, some studies have modelled B-ISDN networks as queueing networks with non-renewal flows. Kroner et al. [28] considered a queueing network in which each connection at the transport level was represented as an on-off source. The long and the short term fluctuations in the links were computed respectively by a fluid flow approximation and by an M/D/1-S model. They showed how to compute the end-to-end delay and the distribution of the transfer delay at each queue in a tandem network. Reising analyzed two queues in tandem [29]. The input of the first queue is a two-state Markov Modulated Bernoulli Process and the interfering traffic of both queues is a Bernoulli process. He derived an approximation model for the output process of the first queue as a two-state MMBP. Grienenfield [30] studied two queues in tandem by using a pertubation method to compute approximations for the mean and the variance of the end-to-end delay too. He considered that both the input and the interfering traffic were wide sense stationary processes, and he used a procedure for matching intermediate results in a Weibuill distribution to compute the jitter delay which is associated with the input traffic (pertubation).

In our investigation, we consider open queueing networks with multiple class of clients. In each node there is a single server with finite buffer space and constant service time. Service is provided in a First-Come-First-Served fashion. Both internal and external traffic are represented as Discrete Time Batch Markovian Arrival processes (figure 1). The discrete time assumption derives from the ATM standard. In order to solve this queueing network with non-renewal flow, we employ the parametric decomposition approximation. The fact that the correlation structure of a cell stream tends to be preserved along the network reinforces the idea of solving the queues in isolation [28]. The network elementary operations are defined as:

The output process of A D-BMAP/D/1/K

The output process of a D-BMAP/D/1/K queue is also correlated, and neglecting its correlations leads to inaccurate results. In [31] the output process of an ATM switch was modelled as two different memoryless processes: i) a Bernoulli process and ii) an Interrupted Poisson Process with the mean duration of the active state equal to the mean busy period. In both cases, the model failed to produce accurate results. At each time slot, at most one cell may depart from a queue. The Markov Modulated Bernoulli Process, a correlated process with single arrivals, is a good candidate for modelling the output process of a D-BMAP/D/1/K queue. We, therefore, developed a procedure for matching the statistics of the output process with the statistics of a two-state MMBP (section IV). Moreover, by modelling the output process as a MMBP, we were able to represent all the flows in the network as D-BMAP process (see figure 2).

Joining

The superposition of two D-BMAP processes with m_1, m_2 states and n_1, n_2 maximum batch size is also a D-BMAP with $m_1 \; x \; m_2$ states and $n_1 + n_2$ maximum batch size. The matrix D_k which elements $(d_{ij})_k$ which give the probability of going from state *i* to state *j* and having a batch arrival of size *k* is computed as:

$$D_k = \sum_{q=0}^{n_1} D_q^{(1)} \otimes D_{k-q}^{(2)}$$

For instance, the superposition of an MMBP and a D-BMAP with maximum batch size of 2 is given by:

$$D_0 = D_0^{(1)} \otimes D_0^{(2)}$$

$$D_1 = D_0^{(1)} \otimes D_1^{(2)} + D_1^{(1)} \otimes D_0^{(2)}$$

$$D_2 = D_0^{(1)} \otimes D_2^{(2)} + D_1^{(1)} \otimes D_1^{(2)} + D_2^{(1)} \otimes D_0^{(2)}$$

$$D_3 = D_0^{(1)} \otimes D_3^{(2)} + D_1^{(1)} \otimes D_2^{(2)} + D_2^{(1)} \otimes D_1^{(2)} + D_3^{(1)} \otimes D_0^{(2)}$$

where $A \otimes B$ denotes the Krockener product of matrix *A* by matrix *B*.

Splitting

We assume that routing is state independent. It means that the probability of a cell departing from one node and going to another node is fixed. When characterizing the flow between two nodes, we represent the output process of the first queue as an MMBP process, and then model the flow that goes to the second queue as an MMBP with parameters:

$$(p_{ij} \times p_1, p_{ij} \times p_2, \alpha_1, \alpha_2)$$

where p_{ij} is the probability that a cell leaves node *i* and goes to node *j*.

IV) THE MATCHING PROCEDURE

Finding an appropriate representation for the output process of a queue is of paramount importance for defining a queueing network framework. With the advent of traffic integration, we now face the challenge of obtaining a suitable representation for the output process of queues with correlated input. Only recently, have some studies derived the statistical properties of the output process of queues with Markov Modulated input [32]-[34]. Saito [32] studied the output process of an N/G/1 queue and particularly of the MMPP/D/1 queue. He analyzed the interdeparture time process and explicitly detailed the expressions for the mean, variance and covariance at lag 1. The burstiness of a process (C_p (z)) was defined as the Z-transform of the covariance of interdeparture time (in [33] C_p (z) was used to analyze integrated traffic). By comparing the C_p (z) curves for the input and for the output processes, Saito concluded that covariances are likely to be preserved. He also pointed out that the reduction of the coefficient of variation is larger in heavily loaded systems than it is in lightly loaded ones. Thus, the departure process of a lightly loaded system with bursty arrival process tends to be bursty too. Takine et al. [34] studied the output process of a D-BMAP/D/1/K queue. They derived not only the expression for the mth-factorial moment of the interdeparture time process, but also the distribution of the idle and the busy periods. They pointed out that, when the variation of arrivals is high, both the mean and the idle periods become larger as the correlation increases. This observation is also valid for the coefficient of variation of busy periods and of the interdeparture time.

Before showing how to match the statistics of the output process with the statistics of a two-state MMBP, we need to characterize the output process itself. Having exactly one departure at each time instants of a busy period suggests that we can represent the output process as a D-MAP in which the matrices D'_1 and D'_0 correspond respectively to busy and idle periods. In order to capture the behavior of busy/idle periods, we need to associate each state of the D-MAP with the phase of the arrival process and with the number of enqueued cells at the end of each time slot [11]. If we have a gated server (i.e., if a cell finds the server empty at its arrival slot, it can only be transmitted at the next slot) then, the output process is given by [11]:

$$D'_0 = \begin{bmatrix} D_0 & D_1 & \dots & D_{k-1} & \sum_{n=k}^{\infty} D_n \\ 0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

$$D'_1 = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ D_0 & D_1 & D_2 & \dots & D_{k-1} & \sum_{n=k}^{\infty} D_n \\ 0 & D_0 & D_1 & \dots & D_{k-2} & \sum_{n=k-1}^{\infty} D_n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & D_0 & \sum_{n=1}^{\infty} D_n \end{bmatrix}$$

On the other hand, if we have a cut-through type of service (i.e. the cell can be transmitted in the same slot in which it arrives) the output process is specified by:

$$D'_0 = \begin{bmatrix} D_0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

$$D'_1 = \begin{bmatrix} D_1 & D_2 & D_3 & \dots & D_K & \sum_{n=K+1}^{\infty} D_n \\ D_0 & D_1 & D_2 & \dots & D_{K-1} & \sum_{n=K}^{\infty} D_n \\ 0 & D_0 & D_1 & \dots & D_{K-2} & \sum_{n=K-1}^{\infty} D_n \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & D_0 & \sum_{n=1}^{\infty} D_n \end{bmatrix}$$

The index of dispersion time curve completely defines the correlation structure of a counting process. Consequently, to accurately approximate the output process, it is important to provide a good match with the index of dispersion time curve. In our procedure, we chose to match the long-term index of dispersion and the covariance of the number of arrivals between consecutive slots and between two consecutive slots (at lags of 1 and 2). Our procedure is:

$$output_{mean} = MMBP_{mean}$$

$$output_{variance} = MMBP_{variance}$$

$$output_{covariance\ lag=1} = MMBP_{covariance\ lag=1}$$

$$output_{covariance\ lag=2} = MMBP_{covariance\ lag\ =\ 2}$$

We showed in [34] that the results produced when the output process of a D-BMAP/D/1/k is substituted by a two-state MMBP are accurate. In the matching procedure validation experiments, the maximum percentual error was below 7% for highly loaded systems and as low as 0.1% for moderated loaded systems.

V) THE END-TO-END DELAY COMPUTATION

To compute the end-to-end delay in an ATM virtual path, we make use of the parametric decomposition approximation. In the parametric decomposition approximation, the queueus are analyzed in isolation only after their input process are fully characterized. In this approach, the dependences among the queues are represented approximately by the flow parameters. We model an *N* node ATM virtual path as *N* queues in sequence (tandem). The input process of the *i^{th}* queue is composed by both the output process of the *i-1^{th}* queue and by an interfering traffic (except for the first queue). We assume that with a fixed probability p_i, a cell may leave the system (figure 3). The computation procedure can be summarized as:

1- For i from1 to N do:

1.1- Characterize the input process by performing a joining operation between the output process of the i-1^{th} queue and the i^{th} interfering process. For the first queue, the input process is given by the input process to the network;

1.2- Compute the steady state queue length distribution of the i^{th} D-BMAP/D/1/K queue. Compute the mean delay seen by an arriving cell at queue i, d_i (see [11], [16] for the solution of the D-BMAP/D/1/K queue);

1.3- Characterize the output process of the i^{th} queue by matching the statistics of the output process with the statistics of a two-state MMBP;

1.4- Characterize the flow that departures from the i^{th} queue and goes to the i+1^{th} queue by a splitting operation;

2- Compute the end-to-end delay $\bar{d} = \sum_{i=1}^{N} d_i$

VI) NUMERICAL EXAMPLES

To validate the computational procedure, we consider two different D-BMAP processes. The first process is a two state D-BMAP with the same transition probability in each state (α) [34]. The batch size is Poisson distributed with mean $(1+c)\rho$ (state 1) and $(1-c)\rho$ (state 2) where ρ is the overall traffic intensity and c is a parameter. It was demonstrated in [34] that the square coefficient of variation (C_v^2) and the correlation coefficient of the number of arrivals at lag n ($C_c(n)$) are respectively given by:

$$C_v^2 = \rho^{-1} + c^2$$

$$C_c(n) = \frac{c^2\rho}{1+c^2\rho} \times (2\alpha-1)^n$$

Unless otherwise stated, the data shown in this section corresponds to a network of five queues in sequence. Each server provides service in a First-Come-First-Served basis and has buffer size of 100. Time is normalized to one slot. The interfering traffic is the same for all queues, except the seeds for the queue random number generator [36]. We show the percentual error of the estimated end-to-end delay which is defined by: $\frac{d_{sim} - d_{est}}{d_{sim}}$ where d_{sim} is the end-to-end delay obtained in a simulation experiment (95% confidence interval) and d_{est} is the end-to-end delay estimated by section VII computational procedure.

Takine et al. [34] pointed out that when the coefficient of variation of the input stream is moderate (c = 0.5) to high (c = 0.9), the correlation of arrivals plays a key role in determining the characteristics of the output process (mean interdeparture time, mean duration of busy period and idle periods) In our experiment, we kept constant two of the three statistics $(\rho, C_v, C_c(1))$ and varied the third one by changing either ρ, c , or α. In figures 4 and 5, we varied C_v by changing c from 0.1 to 0.9 for $\rho = 0.4$ and $\rho = 0.8$, respectively. In order to avoid both the non-queueing phenomenon in tandem queues with constant service time and to make the output process the major responsible for the delay, we kept the interfering traffic intensity low ($\rho = 0.1$, c = 0.1, $\alpha = 0.55$). We verified that the computational procedure captures accurately the variations in C_V. However, we noticed that it performs much better at moderate loads ($\rho = 0.4$, figure 4) than at high loads ($\rho = 0.8$, figure 5). The percentage error was less than 2% at moderate load and was below 8% even at high loads. By varying α from 0.1 to 0.9, we evaluated the impact of both positively and negatively correlated streams in the precision of the computational procedure. No significant impact was observed (figures 5 and 6). The same trend seen in figures 4 and 5 is carried over to figures 6 and 7. The matching procedure gives better delay estimations for lightly loaded systems ($\rho = 0.2$, figure 6) than for highly loaded systems ($\rho = 0.8$, figure 7).

To make sure that the delay estimation is not affected by the buffer size, we fixed both the input and the interfering process, and varied the buffer size of all the queues from 50 to 250. No significant impact was observed. Figure 8 shows an experiment with $\rho = 0.7$, $c = 0.1$ and $\alpha = 0.9$ for both input and interfering processes.

In order to evaluate the accuracy of the computational procedure in a wide range of delay values, we fixed the input process and varied the interfering process. In figure 9, the input process parameters are $\rho = 0.5$, $c = 0.5$ and $\alpha = 0.9$. The interfering process of each queue is the same; c and α are 0.5 and 0.55 respectively, and ρ varies from 0.05 to 0.8. We noticed again that the computational procedure is more accurate for lightly/moderately loaded systems than it is for highly loaded systems. A maximum error of 8.3% was observed.

We analyzed networks of queues in which a cell may leave the system after receiving service. We assumed that the probability of leaving the system (splitting probability) is the same for all queues. In figure 10, we show an experiment whose input process parameters are ($\rho = 0.8$, $c = 0.5$, $\alpha = 0.9$), and the interfering process parameters are ($\rho = 0.1$, $c = 0.1$, $\alpha = 0.55$). The introduction of the splitting operations had no significant impact on the accuracy of the computational procedure. Due to the decrease in the offered load (as observed before) the error decreased as the splitting probability increased.

To ensure that the obtained results were not influenced by the size of the network, we increased the number of queues from 5 to 20. Figure 11 shows an experiment with the same traffic parameters of figure 10. We observed that the number of queues did not affect the accuracy of the estimated values.

The second process used in the validation experiments was a D-BMAP, the parameters of which (transition probability and distribution of batch size) were computed according to a matching procedure which took into account the underload/overload periods of a server. This procedure is quite flexible and one can easily incorporate data, voice and video sources into a stream.

Overall, the maximum observed error of the end-to-end delay estimation was under 8.5%

VI) CONCLUSIONS

In this paper, we introduced a procedure for computing the end-to-end delay in an ATM virtual path. This procedure was shown to be accurate. Errors were less than 8.5% and parameters such as buffer size and network size had no significant impact on the accuracy of the computational procedure. Moreover, we showed a procedure for modelling the output process of a D-BMAP/D/1/K queue, and we described a framework for the analysis of queueing networks with Markov modulated flow. We are currently validating the queueing network framework for more generally connected networks.

VII) REFERENCES

[1] G. Latouche and V. Ramaswami, "A unified stochastic model for the packet stream from periodic sources", *Perfor. Eval.*, 14, pp. 103-121, 1992.

[2] Gihr O. and Tran-Gia P., "A layered description of ATM cell traffic streams and correlation analysis", in *Proc. of IEEE INFOCOM*, pp. 137-144, 1991.

[3] B. E. Helvik, P. Hokstad and N. Stol, "Correlation in ATM traffic streams- some results", in *Proc. of the 13th ITC*, pp. 25-32, 1991.

[4] N. L. S. Fonseca and J. A. Silvester, "Estimating the loss probability in a multiplexer loaded with multi-priority MMPP streams", in *Proc. IEEE ICC*, pp. 1037-1041, 1993.

[5] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet

multiplexer for voice and data", *IEEE J. Select. Areas Commun.*, vol. 4, pp. 833-846, Sep. 1986.

[6] H. Heffes and D. Lucantoni, “A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance”, *IEEE J. Select. Areas Commun.*, vol. 4, pp. 856-868, Sep. 1986.

[7] D. Anick, D. Mitra and M. M. Sondhi, “Stochastic theory of a data-handling system with multiple sources”, *The Bell Sys. Tech J.*, pp. 1871-1894, 1982.

[8] S.Q. Li, “A general technique for discrete queueing analysis of multimedia traffic on ATM”, *IEEE Trans. Commun.*, vol. 39, pp. 1115-1132, Jul.1991.

[9] J. A. S. Monteiro, M. Gerla and L. Fratta, “Statistical multiplexing in AM networks”, *Perfor Eval.*, 12, pp. 157-167, 1991.

[10] N. L. S. Fonseca and J. A. Silvester, “A comparison of push-out policies in an ATM multiplexer”, in *Proc. of IEEE Pac. Rim Conf. on Commun. Comp. and Signal Proc.*, pp. 338-341, 1993.

[11] C. Blondia, “A discrete-time batch Markovian arrival process as B-ISDN traffic model”, *Belgian J. of Oper Res., Stat. and Comp. Science*, vol. 32 (3), pp. 3-23, 1992.

[12] D. M. Lucantoni, “New results on the single server queue with a batch markovian arrival process”, *Stochastic Models*, vol.7, no. 1, pp. 1-46, 1991.

[13] M. F. Neuts, Matrix-geometric solutions in stochastic models: An algorithmic approach, John hopkins University Press, Baltimore, MD, 1981.

[14] M. Neuts, “Models based on Markovian arrival process”, *IEICE Trans. Commun.*, vol E75-B, n0 12, Dec 1992.

[15] S. Wang and J. A. Silvester, “A discrete-time performance model for integrated service ATM multiplexers”, to appear in *Proc. of IEEE GLOBECOM’93,* 1993.

[16] C. Blondia and O. Casals, “Performance analysis of a Statistical multiplexing of VBR sources”, *Proc of IEEE INFOCOM’92,* pp 828-838, 1992.

[17] P. J. Kuehn, “Approximative analysis of general queueing networks by decomposition”, *IEEE Trans. Commun.*, vol COM-27, 1, pp. 113-126, 1979.

[18] W. Whitt, “The Queueing network analyzer”, *The Bell Sys. Tech.l J.*, vol. 62, pp. 2779-2815, Nov. 1983.

[19] O. Hashida, Y. Takahashi and S. Shimogawa, “Switched batch Bernoulli process (SBBP) and the discrete-time SBBP/G/1 queue with application to statistical multiplexer performance”, IEEE J. Select. Areas in Commun, vol. 9, n 3, pp. 394-401, April 1991.

[20] J. J. Bae, T. Suda and R. Simha, “Analysis of individual packet loss in a finite buffer queue with heterogeneous markov modulated arrival processes: a study of traffic bursti-

ness and priority packet discarding", in *Proc INFOCOM'92*, pp. 0219-0230, 1992

[21] Y. Le Boudec, "An efficient solution method for markov models of ATM links with loss priorities", *IEEE J. Select. Areas Commun.*, vol. 9, pp. 408-417, Apr. 1991.

[22] F. Guillemin, J. Boyer and Dupuis, "Burstiness in broadband integrated networks", Performance Evaluation, vol. 15 pp. 163-176, 1992

[23] I. Khan and V. O. K. Li, "Performance analysis at an ATM multiplexer serving a superposition of bursty traffic sources", submitted to publication

[24] F. Baskett, K.M. Chandy, R.R. Muntz and F. Palacios, "Open, closed and mixed networks of queue with different classes of customers", *J. ACM*, pp. 248-260, 1975.

[25] E. de Souza e Silva and R.R. Muntz, "Queueing Networks: Solutions and Applications", in Stochastic Analysis of Computer and Communication Systems, H. Takagi editor, North Holland, 1990.

[26] L. Disney and P. C. Kiessler, Traffic processes in queueing networks: A markov renewal approach, The Johns Hopkins University Press, Baltimore, 1987.

[27] Y. Ohba Y., M. Murata and H. Miyahara , "Analysis of interdeparture processes for bursty traffic in ATM networks", *IEEE J. Select. Areas Commun.*, vol 9, n 3, pp 468-476, April 1991.

[28] H. Kroner, M. Eberspacher, T. H. Theimer, P. J. Kuhn and U. Briem, "Approximate analysis of the end-to-end delay in ATM networks", in *Proc of IEEE INFOCOM'92*, pp. 879-985

[29] J.A. Resing, "ATM cell stream through tandem queues", Mini Symposium on Performance Aspects of ATM Networks, Leidschendam, 1991.

[30] R. Grunenfelder, "A correlation based end-to-end cell queueing delay characterization in an ATM network", in *Proc. of 13th ITC*, vol on queueing and control in ATM, pp. 59-64, 1991.

[31] F. Bonomi, S. Montagna and R. Paglino , "Busy period analysis for an ATM switching element output line", in *Proc. of IEEE INFOCOM*, pp. 544-550, 1992.

[32] H. Saito, "The Departure process of an N/G/1 Queue", *Perfor. Eval.*, 11, pp. 241-251, 1990

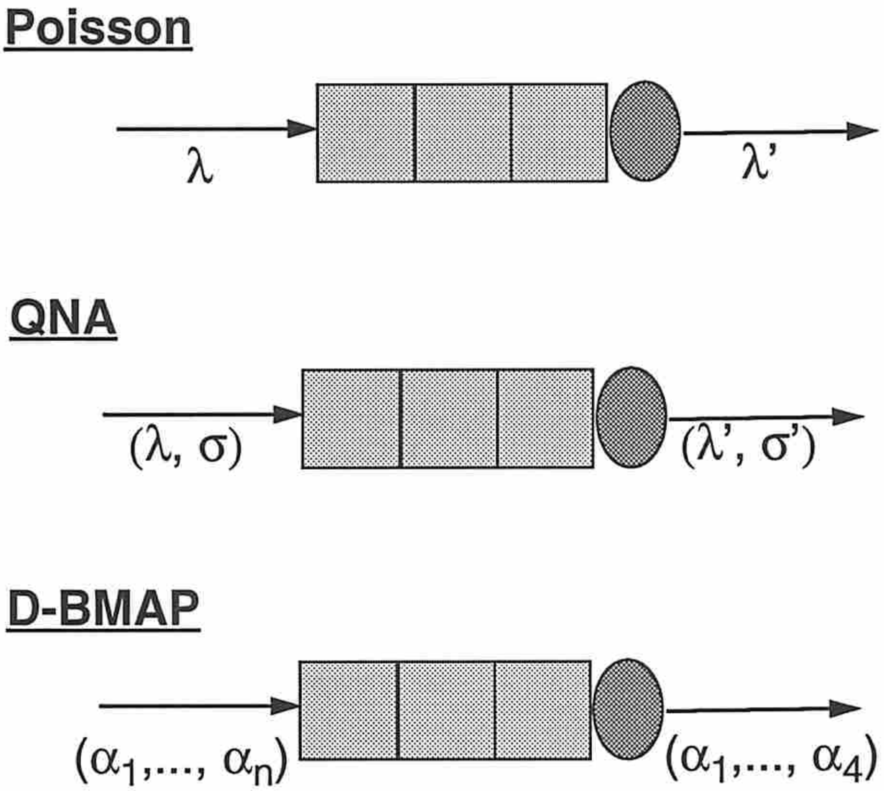
[33] H. Saito, M. Kawarasaki and H. Yamak, "An analysis of statistical multiplexing in an ATM transport network", *IEEE J. Select. Areas Commun.*, vol 9, n 3, pp 359-367, April 1991.

[34] N.L.S. Fonseca and J. A. Silvester, "Modelling the output process of an ATM multiplexer with Markov modulated arrivals", USC CENG Tech Report 93-35, 1993.

[35] T. Takine, T. Suda and T. Hasegawa, "Cell loss and output process analyses of

finite buffer discrete time queueing system with correlated arrivals", in *Proc. of IEEE INFOCOM*, pp 1259-1268, 1993.

[36] S. Lavenberg, Computer Performance Modeling handbook, Academic Press: San Diego, 1983.



Poisson

QNA

(λ, σ)
(λ', σ')

D-BMAP

$(\alpha_1, \ldots, \alpha_n)$
$(\alpha_1, \ldots, \alpha_4)$

Figure 1: In product form networks the flow is characterized by the mean of a Poisson process. In queueing networks with renewal flow by the mean and the variance of the renewal process. In queueing networks with Markov modulated flow, the number of parameters depends on the size of the underlying Markov Chain. If the MC is a two state, we need four parameters.

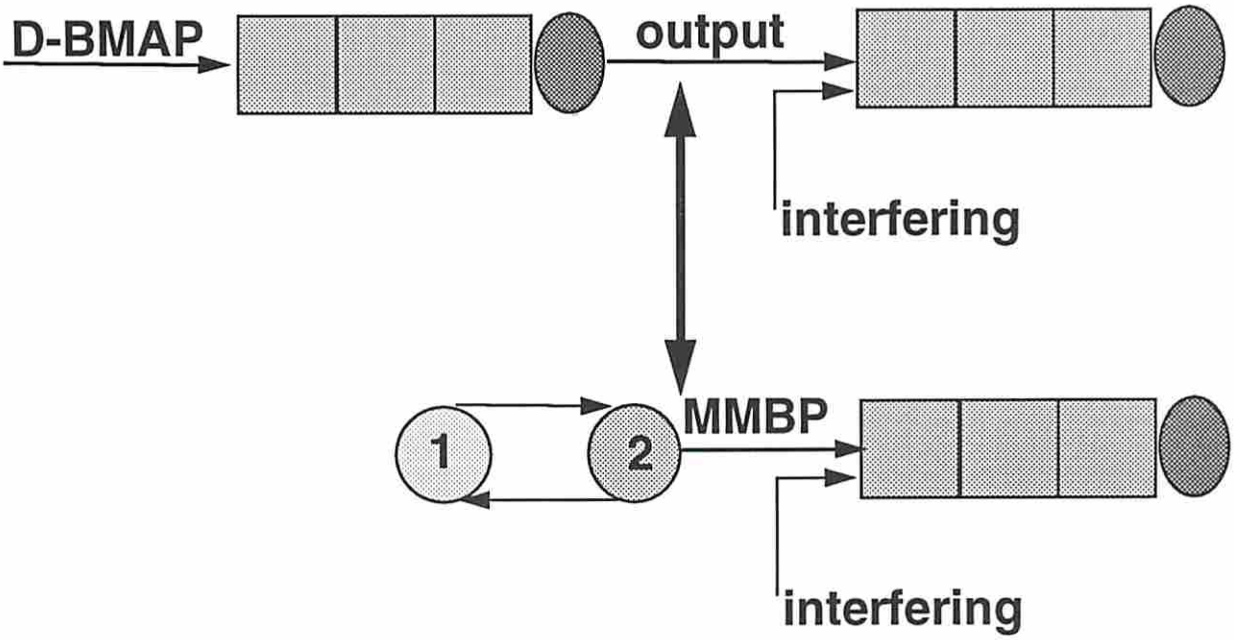


Figure 2: The output process of a D-BMAP/D/1/K is modelled as a two-state MMBP.

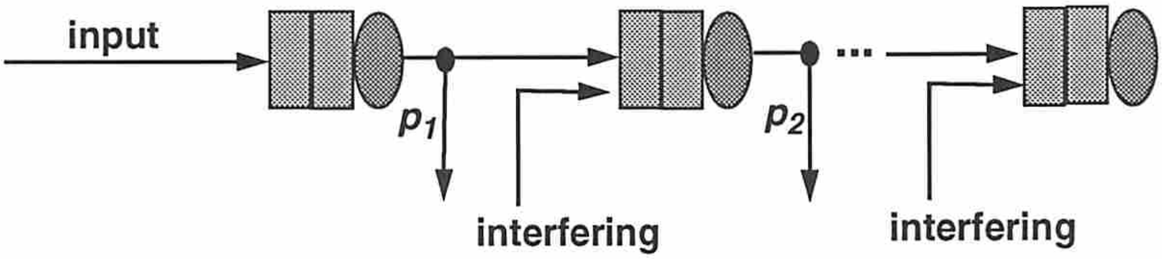


Figure 3: The focused network topology.

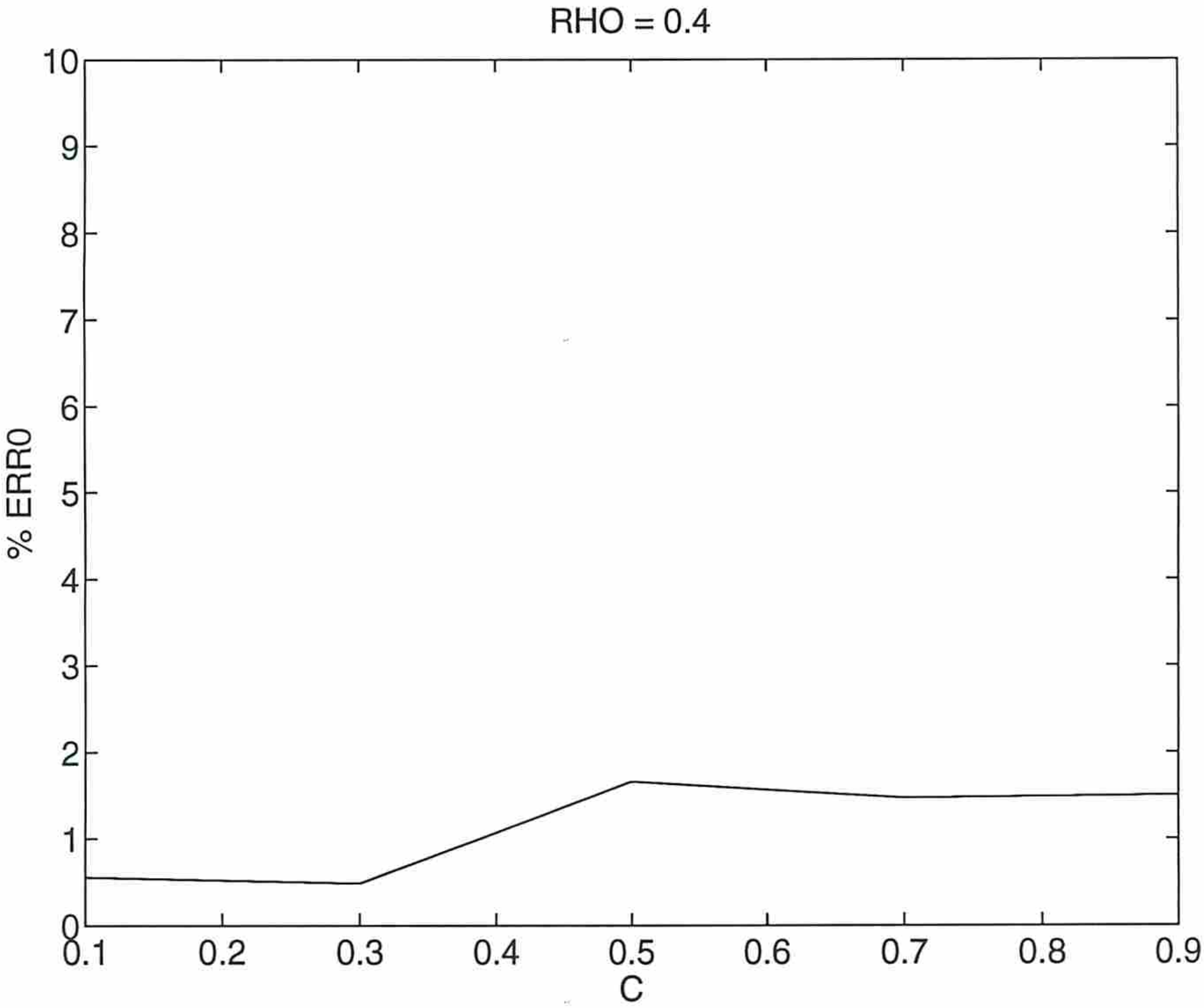


Figure 4: Percentage error as a function of *c* for ρ = 0.4.

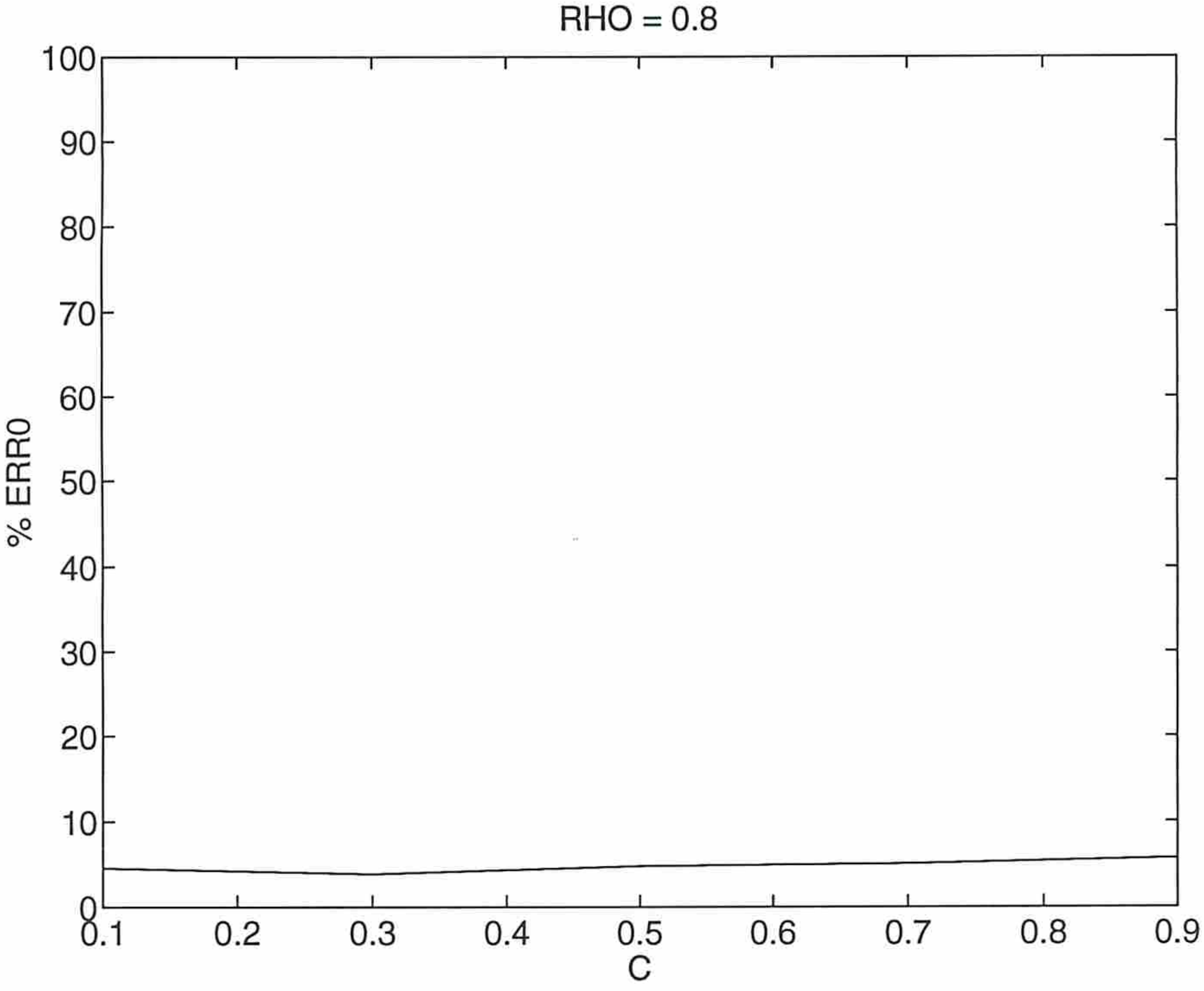


Figure 5: Percentage error as a function of *c* for ρ = 0.8.

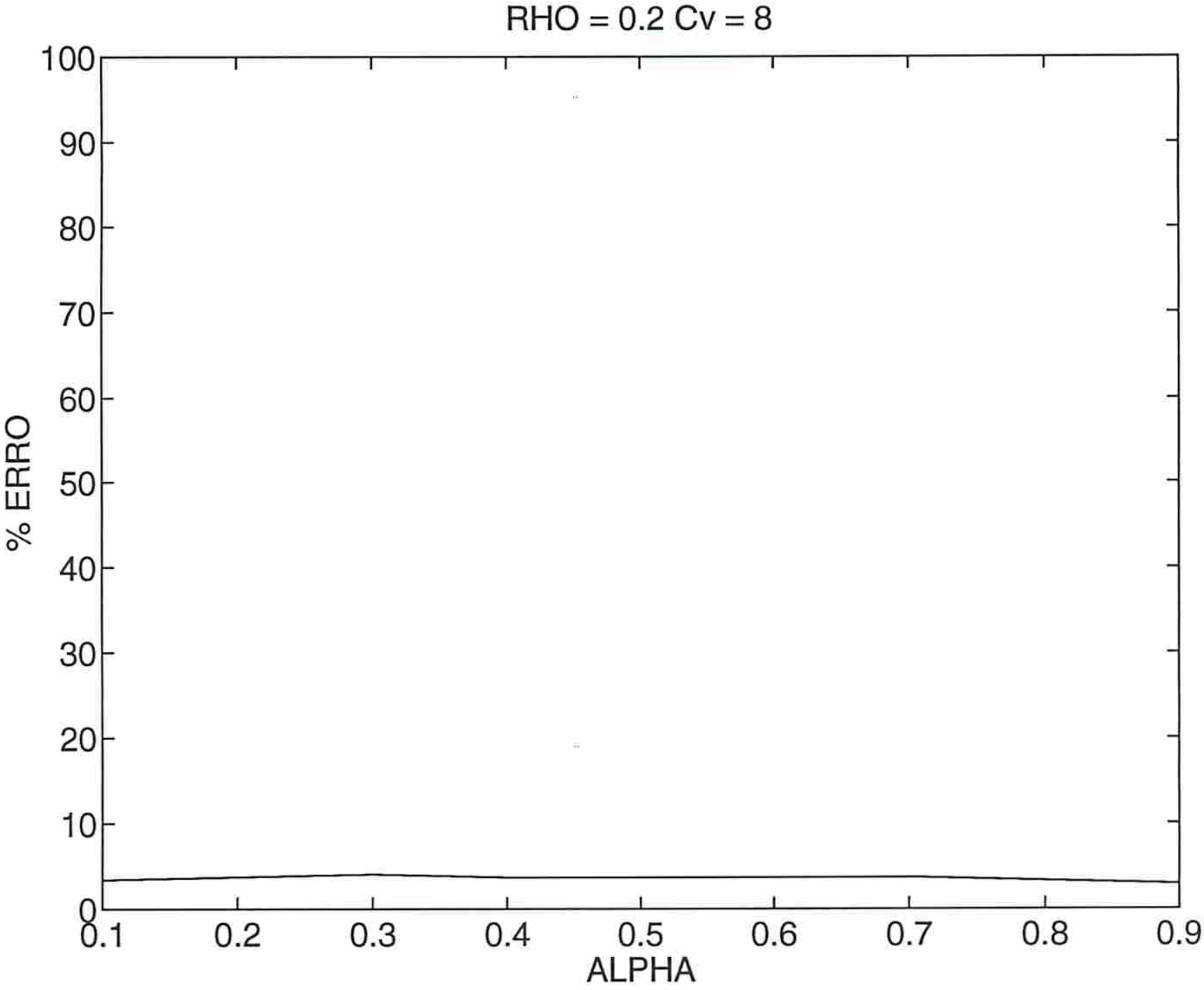


Figure 6: Percentage error as a function of α for ρ = 0.2 and C_v = 8.

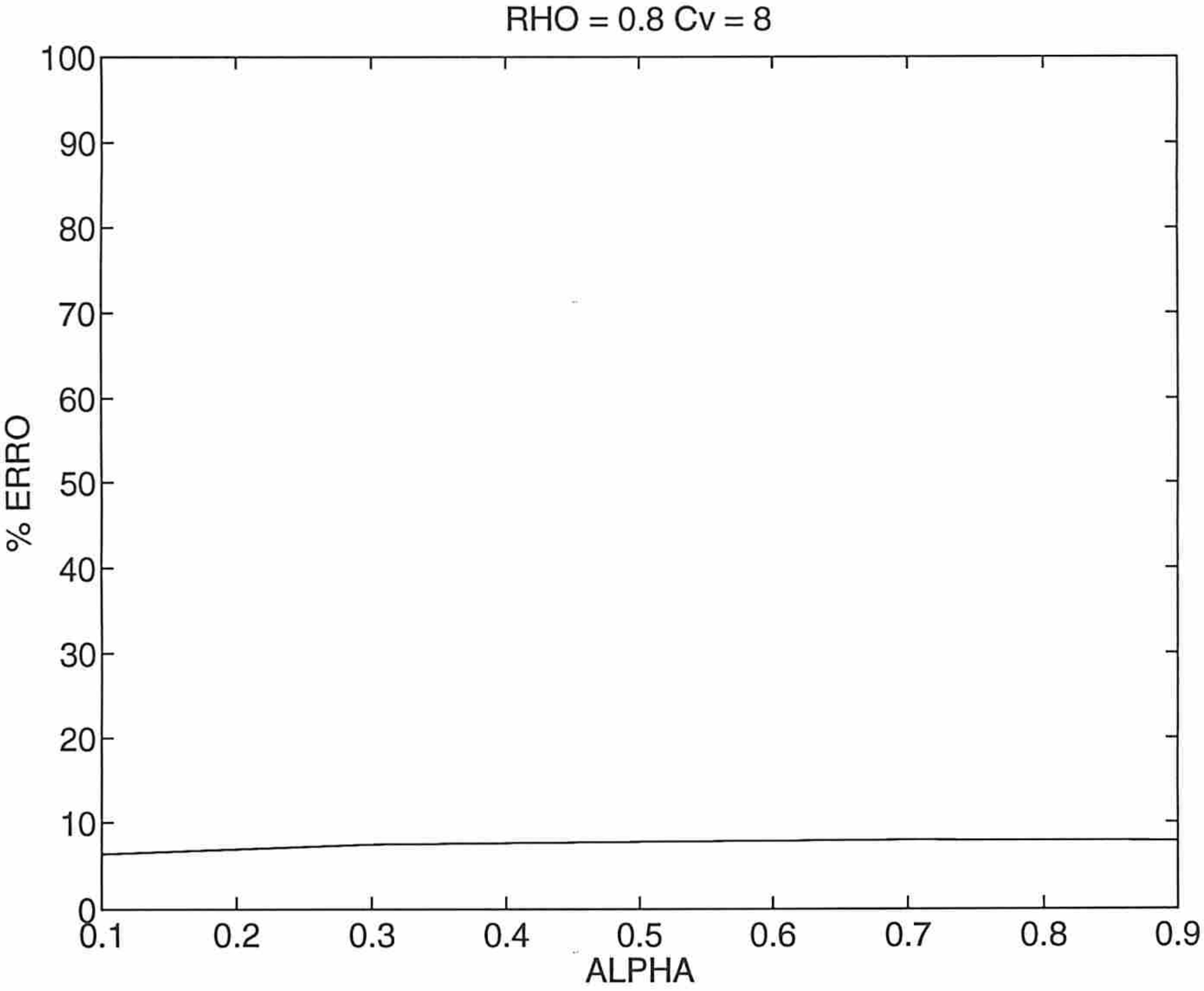


Figure 7: Percentage error as a function of α for ρ = 0.8 and C_v = 8.

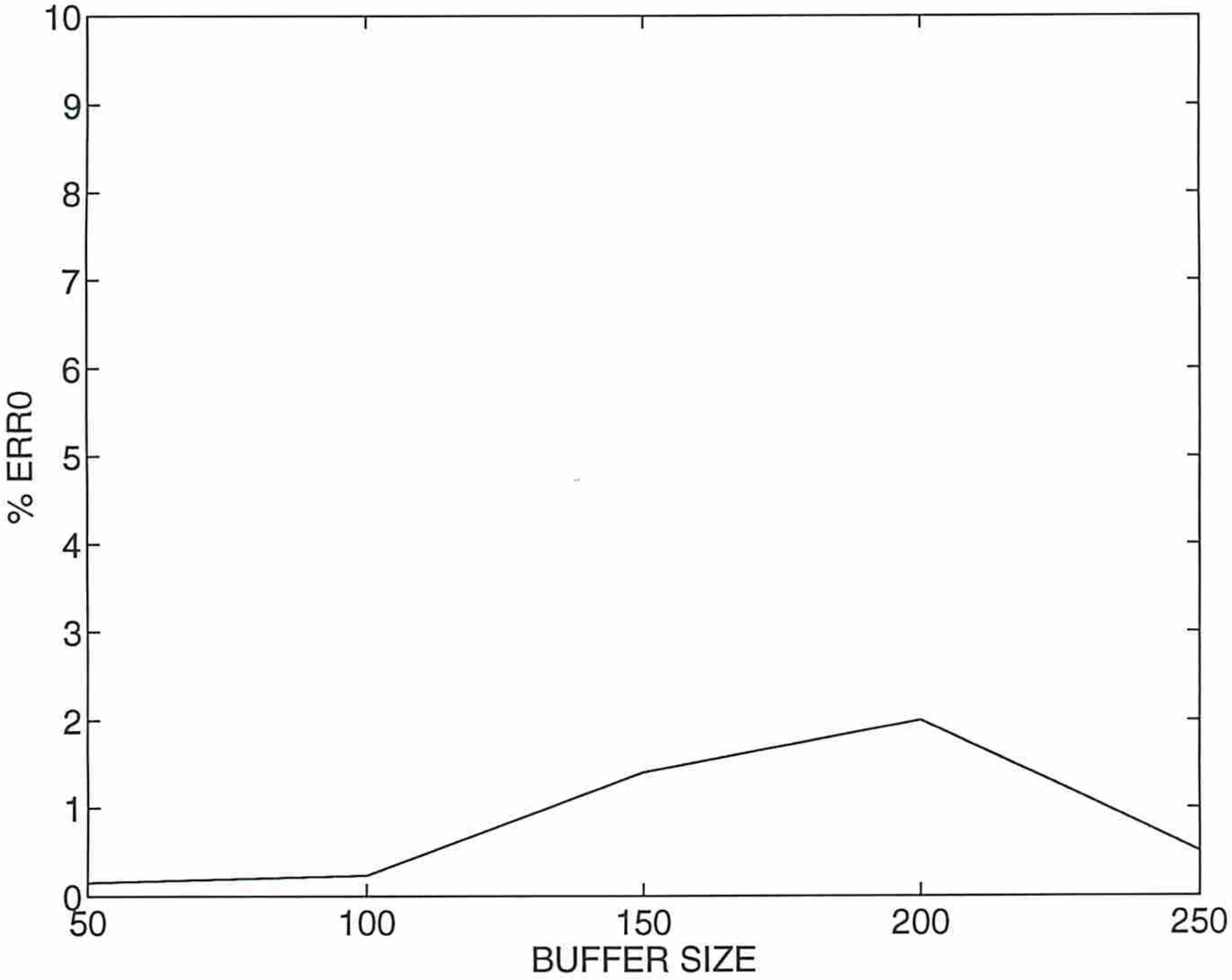


Figure 8: Percentage error as a function of the buffer size.

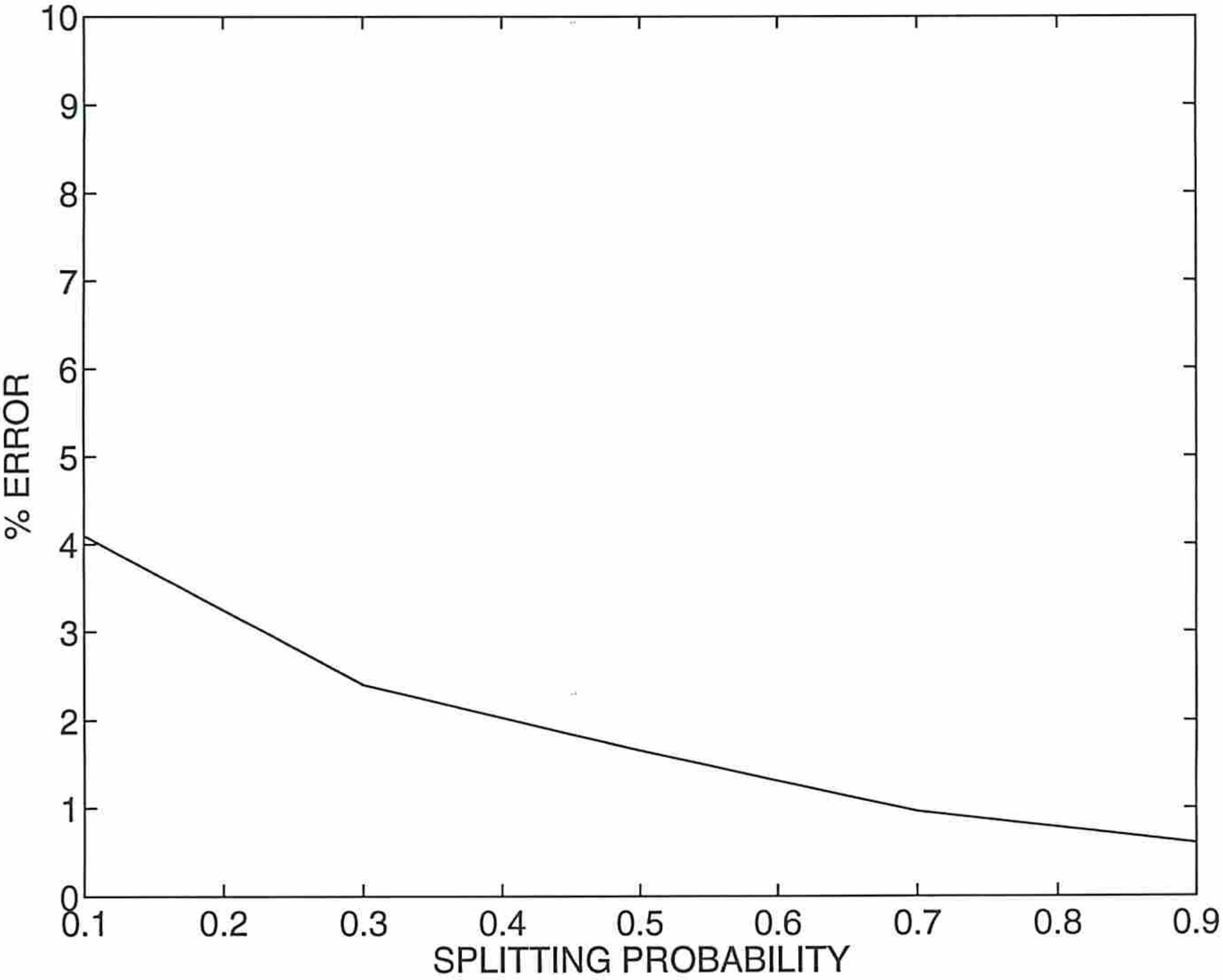


Figure 9: Percentage error as a function of the splitting probability.

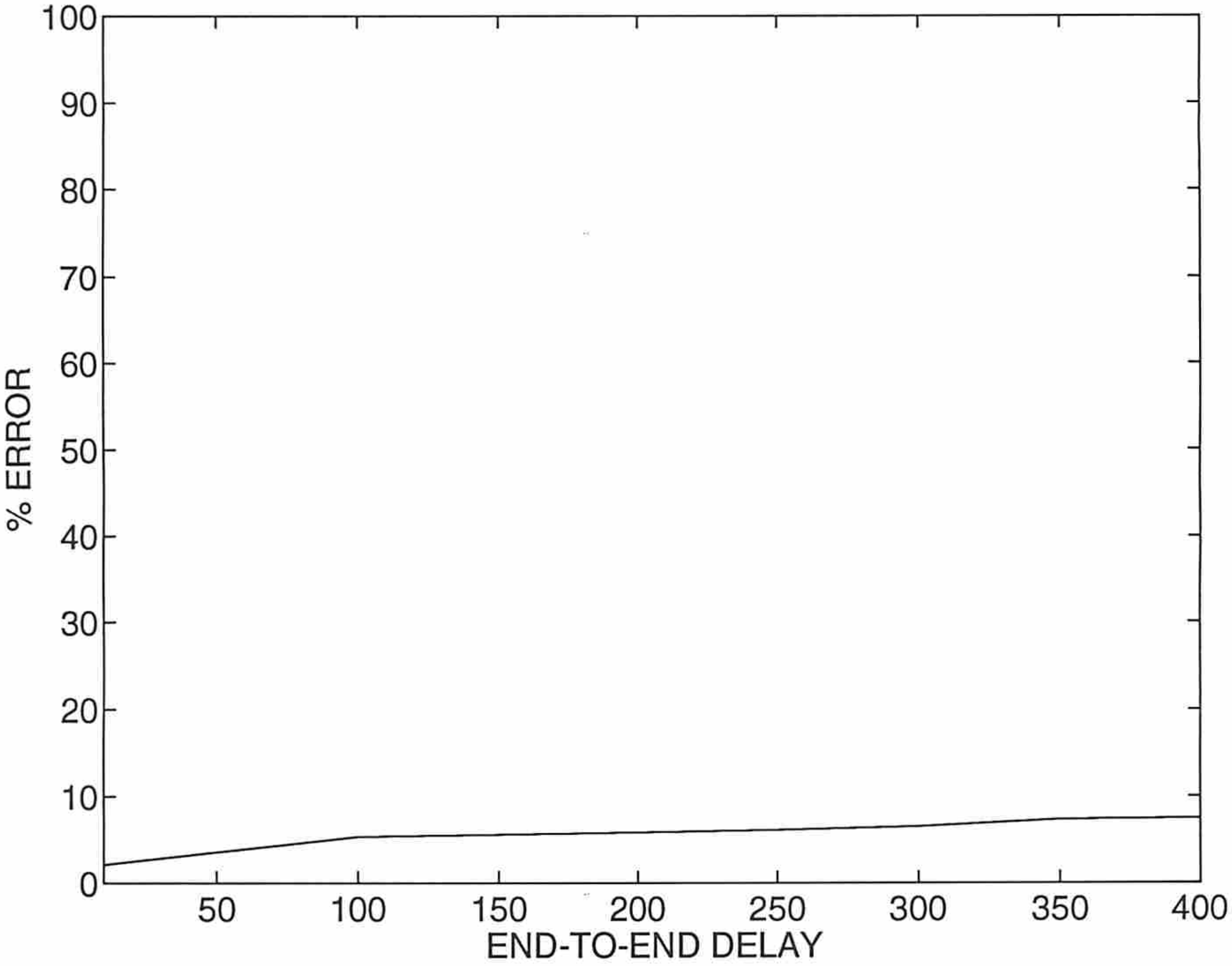


Figure 10: percentage error as a function of the pre-specified end-to-end delay.

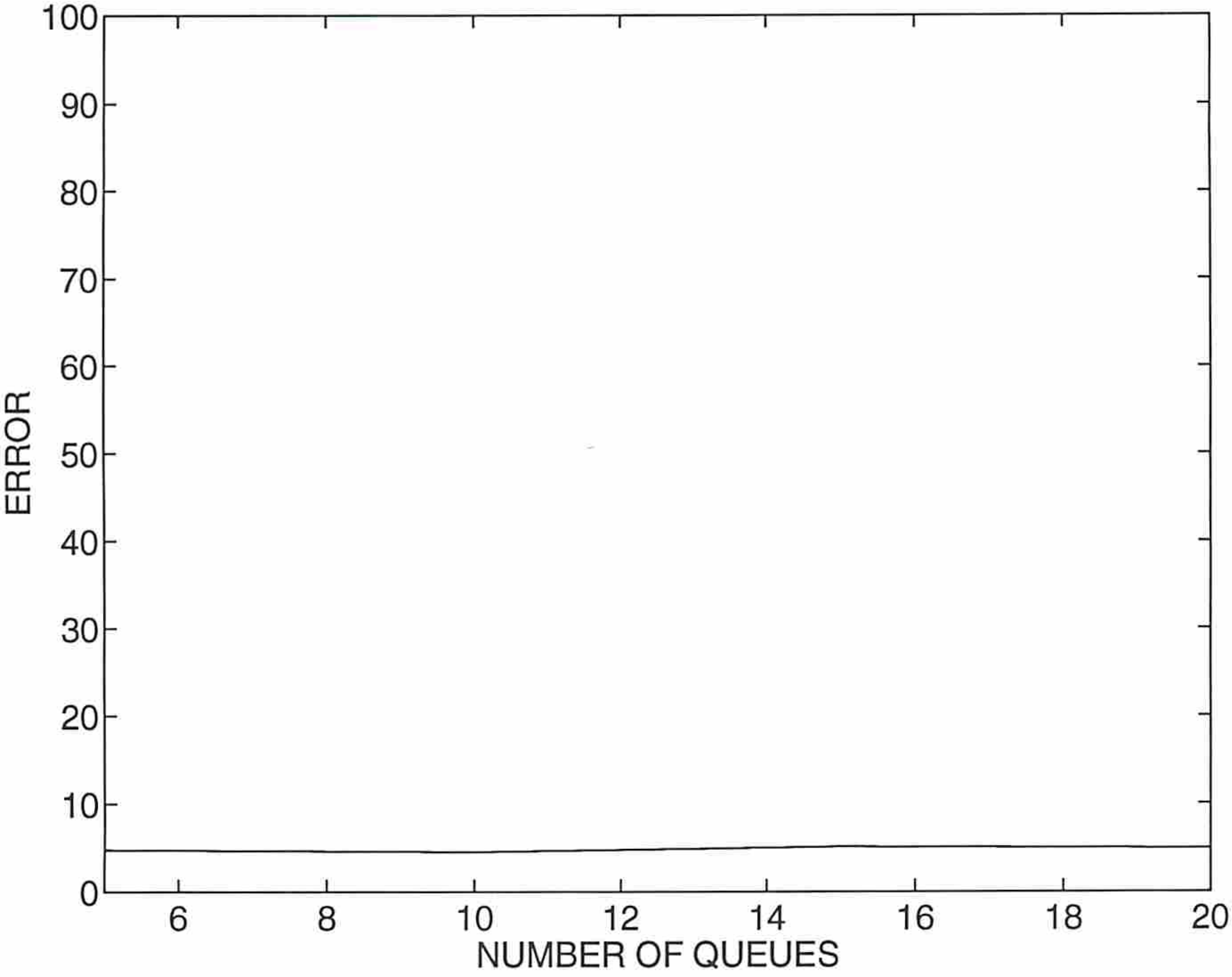


Figure 11: Percentage error as a funtion of the network size.