IP Traffic Prediction and Equivalent Bandwidth for DAMA TDMA Protocols

J. Aracil, D. Morato, E. Magaña, M. Izal
Universidad Pública de Navarra, 31006 Pamplona, SPAIN
email:javier.aracil@unavarra.es

***Abstract*— The use of IP traffic prediction techniques for DAMA TDMA protocols is investigated in this paper. The predicted traffic *distribution* is derived when the input traffic shows long-range dependence features. Furthermore, an equivalent bandwidth is calculated, which allows the wireless terminal to request a certain amount of bandwidth (slot duration) in terms of a target traffic loss probability. The numerical results indicate very good traffic prediction capabilities, together with moderate bandwidth loss.**

I. Introduction and problem statement

Wireless networks, both terrestrial and satellite based, are expected to carry a large fraction of Internet traffic. For all such networks, the multiple access scheme plays a crucial role in providing efficient utilization of the wireless bandwidth. Precisely, due to the traffic burstiness, fixed assignment multiple access techniques are not as cost-effective as Demand-Assignment Multiple Access (DAMA) techniques in the Internet scenario. Such DAMA techniques are normally based on TDMA (or MFTDMA) with variable slot length, in such a way that data bursts can be accommodated on-demand. Due to the protocol felxibility in bandwidth allocation, a wide variety of wireless systems use DAMA-TDMA as the access protocol. For example, DAMA-TDMA techniques are frequently used in satellite-based LAN interconnection networks [1], [2].

Despite of the many variants of DAMA protocols, the following approach can be adopted for modeling purposes [3], [4]. A reservation request message from the source is released per frame or group of frames and the corresponding bandwidth allotment (slot duration) is sent in response from the bandwidth scheduler, which may be located at the master station or on-board in case of OBP satellites. The fundamental issue, however, is to be able to accurately calculate the amount of bandwidth to be requested. Since the propagation delay to the scheduler may be non-negligible (as happens in satellite networks), resources must be reserved not only for the current backlog, but also for the traffic arriving during the time interval elapsed between the release of the bandwidth allocation message and the arrival of the response from the scheduler. In what follows, let us consider that time is slotted in RTT-slots. We will assume that sources have the chance to produce a reservation request once per RTT to the scheduler, i.e. a new bandwidth allocation request will not be sent before the response from the previous bandwidth allocation message has been received[1]. Note that the above scenario does not preclude that the frame duration may be less than a Round-Trip Time (RTT). This is usual for satellite networks, for instance. However, the number of reservation slots per source is reduced to one per RTT. This approach allows to decrease the number of signaling mini-slots per station, in comparison to frame-by-frame allocation techniques. As a result, the control part of the upstream frame is shortened and the transmission efficiency is increased. On the other hand, the upstream frames are reconfigured (due to bandwidth allocations) only once per RTT, thus simplifying network control and synchronization.

With this assumptions in mind, figure 1 shows a reference model for DAMA TDMA protocols. Time is slotted in RTT-slots and, at the beginning of the current RTT-slot k, a reservation request is sent from the user station through reservation mini-slots, possibly involving contention among the stations. Then, a bandwidth allocation message is received in response from the bandwidth scheduler, with the allocated bandwidth for the next RTT-slot $k+1$ (figure 1). Since only traffic already backlogged at the beginning of RTT-slot k is covered by the reservation request message, traffic arriving *during* RTT-slot k is necessarily buffered until transmission in RTT-slot $k+2$. Actually, the reservation request message that includes traffic arriving during RTT-slot k is sent at the beginning of RTT-slot $k+1$. Thus, in order to reduce access latency and increase channel utilization, the bandwidth allocation for the next RTT-slot should also include resources for traffic arriving during the current RTT-slot.

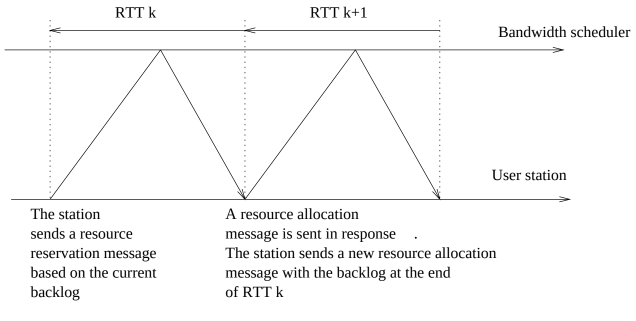
RTT k
RTT k+1
Bandwidth scheduler
User station
The station sends a resource reservation message based on the current backlog
A resource allocation message is sent in response .
The station sends a new resource allocation message with the backlog at the end of RTT k

Fig. 1. DAMA TDMA frame

As a first approximation , a Linear Minimum Square Error Estimate (LMSEE) can be used to estimate traffic arriving during the current RTT-slot. The use of a traffic prediction technique in the DAMA TDMA scenario depicted in figure

[1] Processing time at the scheduler is neglected

1 is advantageous for many-fold reasons: i) it allows to decrease the access delay by sending the reservation message in advance, i.e, before the traffic has been queued at the wireless station, ii) it facilitates flow control from the scheduler to the stations since the control loop between scheduler and station is shortened[2] and iii) it allows extra-time for the scheduler to perform bandwidth allocation since a traffic prediction is available well in advance. In this paper, we go one step further and calculate an *equivalent bandwidth estimate*, which, based on a loss rate objective, provides the bandwidth allocation to be requested for the next RTT-slot. In order to achieve such equivalent bandwidth calculation, one needs to consider the a-priori *distribution* of the number of bytes arriving at the current RTT-slot, instead of the LMSEE solely. The basic assumption is that the incoming traffic follows a stationary Gaussian process[3] with long-range dependence features. In order to understand how traffic correlation can be exploited to provide accurate estimates of the incoming traffic let us briefly present the concept of long range dependence.

A. Traffic self-similarity

Let $\{Z(t), t \in R\}$ be the continuous time process of number of bytes transmitted in the interval $[0, t)$ and consider the discrete-time process $\{X_k = Z(k\delta) - Z((k-1)\delta), k \in N, k \geq 1\}$, being δ a measurement interval. Note that X denotes the (stationary) discrete process of number of bytes per time interval δ. Now, consider the *aggregated* process

$$X_i^{(n)} = \frac{1}{n} \sum_{k=n(i-1)}^{ni} X_k, \qquad n > 1, i \geq 1 \tag{1}$$

and let $\rho^{(n)}(j)$ with $j > 1$ be the autocorrelation function of $\{X_i^{(n)}, i \geq 1\}$. The process $\{X_k, k \geq 1\}$ is *asymptotically second-order self-similar* if

$$\lim_{n\to\infty} \rho^{(n)}(j) = \frac{1}{2}((j+1)^{2H} - 2j^{2H} + (j-1)^{2H}) \tag{2}$$

where H is the Hurst (or self-similarity) parameter. For $1/2 < H < 1$ the autocorrelation function in equation 2 decays slowly, thus being not summable, and we call X_k *long-range dependent*. Note that n in equation (1) defines a traffic timescale. On the other hand, equation (2) states that self-similarity is an asymptotic property, namely, it only happens when $n \to \infty$. In practice, there is a cutoff timescale (δ) beyond which the traffic behaves as a stationary Gaussian self-similar process with constant H parameter [6], while the short timescales show complex, multifractal behavior. This behavior has been clearly identified in a number of recent studies [7] that confirm that there is no single characterization for traffic at all timescales. Intuitively, the number of packets per interval can be arbitrarily small if we select a timescale small enough. Hence, for a very short timescale the marginal distribution of the arrival process is not Gaussian but discrete. As we increase the timescale, by the Central Limit Theorem, the statistical multiplexing of packets coming from a larger number of sources results in a Gaussian process. On the other hand, as the network bandwidth increases more packets from different sources can be accommodated in smaller timescales. Thus, for timescales beyond a cutoff value the number of bytes per interval are well modeled by a Fractional Gaussian Noise (FGN)[4].

As a conclusion, for packet-switched networks the traffic dynamics at low timescales are relevant, specially at low or intermediate load [8]. However, in our case study (TDM frames), we are concerned with *the number of bytes per RTT-slot only*. Since a large number of packets can be accommodated in a RTT-slot (which may be large in satellite networks) we may safely assume that the *number of bytes per RTT-slot* can be characterized as a FGN.

B. Contribution

In this paper we provide a traffic prediction scheme for use in DAMA TDMA systems which is based in a Fractional Gaussian Noise, a commonly accepted model for traffic in a LAN. Our results provide the *distribution* of the number of incoming bytes in future frames and not only a predictor in the minimum square error sense. Consequently, we provide an expression for *the equivalent bandwidth to be allocated in order to meet a traffic loss rate objective*. Both numerical and simulation results asses the performance of the equivalent bandwidth estimator concerning loss rate and link utilization. The rest of the paper is organized as follows: in section II we present the analysis, followed by the results and discussion in section III. Finally, we present the conclusions that can be drawn from this paper.

II. Analysis

First, let us consider that time is slotted in RTT-slots of duration δ. The incoming traffic is defined as the increments of a Fractional Brownian Motion $\{A_t = \mu t + \sigma Z_t, t > 0\}$ being $\{Z_t, t > 0\}$ a standard FBM. Thus, the incoming traffic is a FGN $\{X_n, n \in N\}$ with $X_n = A_{n\delta} - A_{(n-1)\delta}, n > 0$. We note that X_i denotes the number of bytes which are received at the satellite terminal during RTT-slot i. We wish to estimate X_i with the information provided by $X_{i-1}, \ldots, X_{i-n}$. Since the process is stationary the problem is equivalent to finding a distribution for X_{n+1} provided that $X_1, \ldots, X_n$ are known. Since $\{X_n, n > 0\}$ is a Gaussian process any finite set of the random variables X_n's is a multivariate Gaussian random variable with mean μ and covariance matrix $\mathbf{\Sigma} = \{S_{ij}\}$. For a FGN the covariance matrix of the multivariate Gaussian variable $(X_1, \ldots, X_{n+1})$ is defined as follows:

$$S_{ij} = \frac{1}{2}\sigma^2 \left[(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\right] \tag{3}$$

[2] This is specially interesting for ABR services in ATM over satellite [5]

[3] The process is assumed to be truncated to the positive values

[4] An FGN is defined as the increments of a Fractional Brownian Motion [6].

with $i, j = 1, \ldots n+1$. Let us define $T(x_1, \ldots, x_n)$ as the random variable X_{n+1} conditioned to $(X_1 = x_1, \ldots, X_n = x_n)$, namely $T(x_1, \ldots, x_n) = X_{n+1}|(X_1 = x_1, \ldots, X_n = x_n)$. Then, T is a normal random variable $N(\mu^*, \sigma^*)$ with mean and variance [9, Theorem 3.3.1]

$$\mu^* = \mu + \mathbf{\Psi_{21}}\mathbf{\Psi_{11}}^{-1}(x_1 - \mu, \ldots, x_n - \mu)' \tag{4}$$

$$(\sigma^*)^2 = \sigma^2 - \mathbf{\Psi_{21}}\mathbf{\Psi_{11}}^{-1}\mathbf{\Psi_{12}} \tag{5}$$

where $\mathbf{\Psi_{21}} = (S_{(n+1)1}, \ldots, S_{(n+1)n})$, $\mathbf{\Psi_{12}} = \mathbf{\Psi_{21}}'$ and $\mathbf{\Psi_{11}} = (S_{ij})$ for $i, j = 1, \ldots n$. Note that (4) and (5) provide the parameters for the a-priori distribution of the number of bytes arriving at RTT-slot $n+1$, provided that the number of bytes arriving at previous RTT-slots $1, \ldots, n$ are known. In what follows, the number of past RTT-slots to be used in the prediction (n) is set to 5. In accordance to previously published studies [10] only samples in the recent past matter, with the rest of the samples having a slight incremental value in the prediction. We have also verified this hypothesis in a previous paper using 5 samples only [11].

A. *Equivalent bandwidth*

In this section we provide an expression for an equivalent bandwidth $C_{\epsilon,\alpha}$, which is measured in bytes per RTT-slot, with the form

$$P(LR(C_{\epsilon,\alpha}) \leq \epsilon) > \alpha \tag{6}$$

being LR the traffic loss rate (bytes) and ϵ, α two real parameters such that $\epsilon > 0, 0 < \alpha < 1$. If the wireless link provides ATM service then the loss rate can be translated immediately to Cell Loss Rate (CLR). Therefore, the above definition is in accordance with the ATM standards [12]. Furthermore, the above equivalent bandwidth definition offers scope for *statistical* QoS guarantees.

Since the channel is slotted in RTT-slots we focus on the LR within a RTT-slot and impose the condition defined in (6) slot by slot. Recall that X_i represents the number of bytes transmitted in RTT-slot $i, i \geq 1$.

Conditioned to $(X_1 = x_1, \ldots, X_n = x_n)$, X_{n+1} has a normal distribution $N(\mu^*, \sigma^*)$ with μ^* and σ^* defined by equations 4 and 5. Thus, for all $\epsilon > 0$,

$$P(LR \leq \epsilon) = P\left(0 < \frac{X_{n+1} - C_{\epsilon,\alpha}}{X_{n+1}} \leq \epsilon\right) \tag{7}$$

and, consequently, equation 6 is fulfilled if

$$P(LR \leq \epsilon) = P\left(X_{n+1} \leq \frac{C_{\epsilon,\alpha}}{1-\epsilon}\right) > \alpha \tag{8}$$

for all n. Let F_Z be the standard normal distribution function, then, for each RTT-slot, the equivalent bandwidth is defined as the real number $C_{\epsilon,\alpha}$ that fulfills

$$F_Z\left(\frac{\frac{C_{\epsilon,\alpha}}{1-\epsilon} - \mu^*}{\sigma^*}\right) > \alpha \tag{9}$$

An explicit upper bound for the equivalent bandwidth can be provided using the approximation for the residual distribution function of the standard Gaussian distribution $\phi(y) \sim exp(-y^2/2)$ as follows

$$exp\left\{-\frac{1}{2}\left(\frac{\frac{C_{\epsilon,\alpha}}{1-\epsilon} - \mu^*}{\sigma^*}\right)^2\right\} < 1 - \alpha \tag{10}$$

and, finally,

$$C_{\epsilon,\alpha} \geq (1-\epsilon)\left[\mu^* + \sigma^*\sqrt{Log\frac{1}{(1-\alpha)^2}}\right] \tag{11}$$

The last equation is an approximate expression for the equivalent bandwidth per RTT-slot, which is tighter as ϵ increases. The exact equivalent bandwidth can be derived from (9) at nearly no computational cost (see [13, section 13.4] for approximations of a Gaussian distribution function). We must emphasize that $C_{\epsilon,\alpha}$ provides number of bytes to be allocated in the next RTT-slot and changes slot by slot since it depends on the tuple (μ^*, σ^*) which in turn depends on the traffic samples $x_1, \ldots, x_n$ (see (4) and (5)).

III. Results and discussion

Numerical results from the equivalent bandwidth equations derived in the previous section -approximate (11) and exact (9)- are presented in this section. On the other hand, simulation experiments have also been performed. Traffic sources are assumed to carry traffic from a multiplex of end users and their traffic parameters are set to those from the *Bellcore traces* (coefficient of variation $c_v^2 = \sigma^2/\mu^2 = 0.1$, Hurst parameter $H = 0.78$), which have also been used in other studies [6], [14], [10]. We note that loss is due to bandwidth under-provisioning for the incoming traffic. Such overflow traffic may be either discarded or buffered for transmission in subsequent RTT-slots. Thus, the loss rate can be interpreted as the percentage of overflow bytes.

A. *Dynamic behavior*

First, we evaluate the *dynamic* behavior of the equivalent bandwidth allocation, namely, how closely does the equivalent bandwidth follow the original traffic per RTT-slot. We set $\alpha = 0.9$ and provide numerical results for large loss ($\epsilon = 0.2$) and small loss ($\epsilon = 0.01$) in figures 2 and 3. The figures show a realization of the real traffic, the allocated bandwidth per RTT-slot and the traffic loss (overflow bytes) for 200 RTT-slots. The y-axis represents bytes per slot and the x-axis slot number. For both figures 2 and 3, the top plot refers to the exact equivalent bandwidth (9) whereas the bottom plot refers to the approximate equivalent bandwidth (11).

Overall, we observe that the equivalent bandwidth derived in the previous section provides good performance both in terms of loss probability and bandwidth utilization. However, the approximate equivalent bandwidth (11) tends to overestimate resources and it is more accurate as the loss rate objective increases. The exact equivalent bandwidth, on the other hand, follows closely the original traffic.

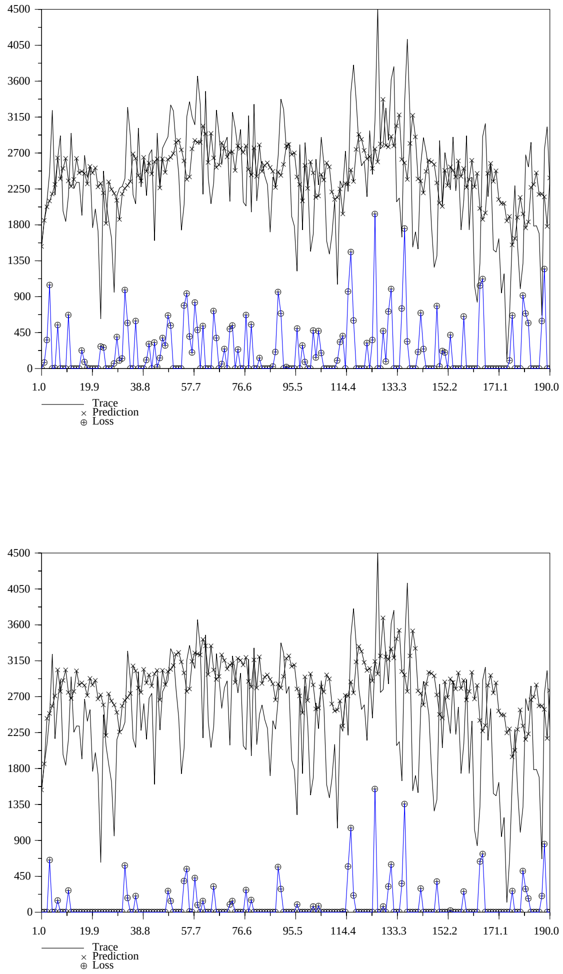

Fig. 2. Real traffic versus prediction and losses ($\epsilon = 0.2, \alpha = 0.9$) -Exact (top) and Approximate (bottom)-

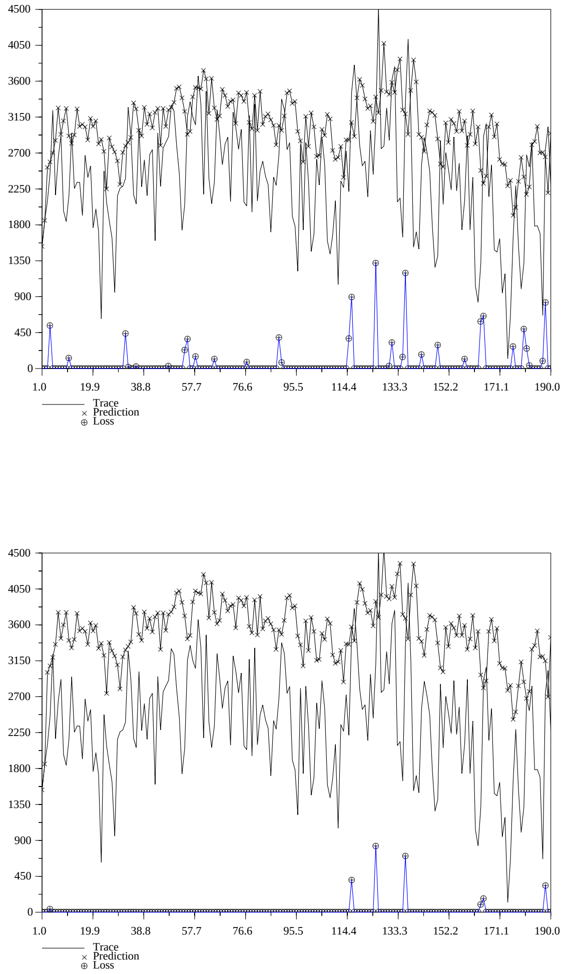

Fig. 3. Real traffic versus prediction and losses ($\epsilon = 0.01, \alpha = 0.9$) -Exact (top) and Approximate (bottom)-

B. *Equivalent bandwidth*

In order to verify that the equivalent bandwidth expressions are accurate, both in the approximate version (11) and in the exact version (9), we run simulation experiments in order to assess that (11) and (9) fulfill (6). Namely, given a loss rate objective (ϵ) and a probability objective (α), we wish to verify that $P(LR \leq \epsilon) > \alpha$.

Figure 4 -top- shows $P(LR \leq \epsilon)$ for $\alpha = 0.9$ (y-axis) versus values of loss probability ϵ (x-axis) in the range $(0.001, 0.2)$. On the other hand, the link utilization factor is shown in figure 4 -bottom-. The y-axis shows utilization factor and the x-axis shows loss probability ϵ in the range $(0.001, 0.2)$

The results show that the approximate equivalent bandwidth expression (11) provides better quality of service (larger α) but at the expense of lower utilization. This is in accordance with the results presented in the previous section, which showed that the approximate equivalent bandwidth expression tends to provide more bandwidth that the exact counterpart (figures 2 and 3).

On the other hand, figure 5 also shows $P(LR \leq \epsilon)$ -top- and link utilization -bottom- for $\alpha = 0.8$. Overall, simulation results show very good agreement with the analytical expressions. Finally, it must also be noted that since analytical expressions for the equivalent bandwidth are provided these experiments are easily reproducible.

IV. Conclusions

An equivalent bandwidth for DAMA-TDMA systems has been provided, that allows to explicitly calculate the requested bandwidth in terms of a traffic loss probability objective *before* the traffic has arrived at the source. The numerical and simulation results show that a loss rate objective is guaranteed, with a given probability objective. Therefore, the equivalent bandwidth expressions can be used to effectively decrease access delay in a wide variety of DAMA-TDMA settings.

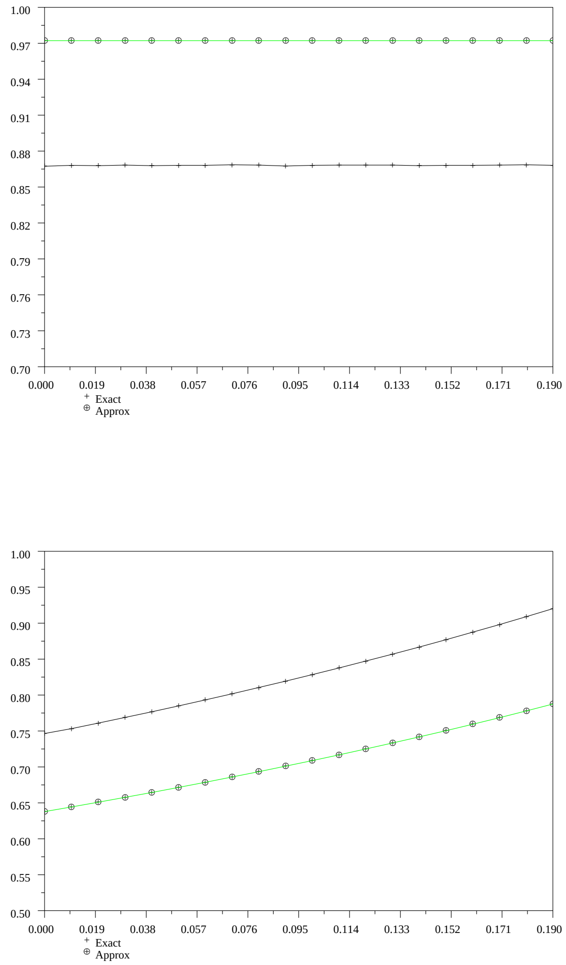

Fig. 4. $P(LR \leq \epsilon)$ -top- and utilization factor -bottom- versus loss rate $(\alpha = 0.9)$

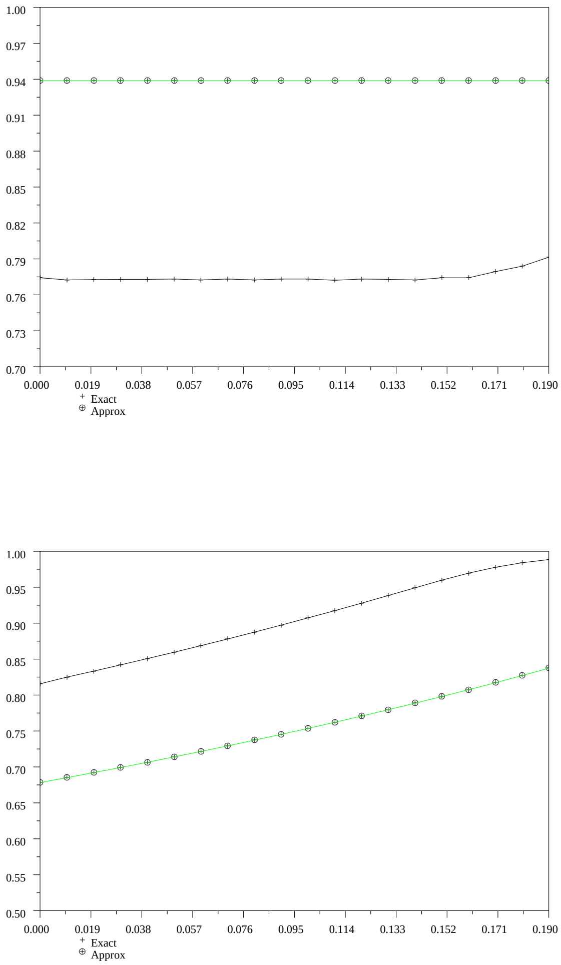

Fig. 5. $P(LR \leq \epsilon)$ -top- and utilization factor -bottom- versus loss rate $(\alpha = 0.8)$

References

[1] Y. F. Hu, G. Maral, and E. F. (Editors), *Service Efficient Network Interconnection via Satellite-EU COST action 253*. John Wiley and Sons, 2002.

[2] F. J. Ruiz, A. Fernandez, C. Miguel, J. Aracil, L. Vidaller, and J. Perez, “The picoterminal network: Portable communications via satellite,” in *TERENA Joint European Networking Congress*, Tel Aviv, Israel, May 1995.

[3] S. Biswas and R. Izmailov, “Design of a fair bandwidth allocation policy for VBR traffic in ATM networks,” *IEEE/ACM Transactions on Networking*, vol. 8, no. 2, pp. 212–223, 2000.

[4] Z. Jiang, Y. Li, and V. C. Leung, “A predictive demand assignment multiple access protocol for broadband satellite networks supporting internet applications,” in *IEEE ICC*, New York, NY, June 2002.

[5] Z. Sun, T. Ors, and B. Evans, “Interconnection of broadband islands via satellite – experiments on the RACE II CATALYST project,” in *Transport Protocols for High-Speed Broadband Networks, Workshop at IEEE GLOBECOM ’96*, 1996.

[6] I. Norros, “On the use of Fractional Brownian Motion in the theory of connectionless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 953–962, August 1995.

[7] A. Feldmann, A. C. Gilbert, and W. Willinger, “Data networks as cascades: Investigating the multifractal nature of internet WAN traffic,” in *SIGCOMM*, 1998, pp. 42–55. [Online]. Available: citeseer.nj.nec.com/feldmann98data.html

[8] A. Erramilli, O. Narayan, A. Neidhart, and I. Saniee, “Performance impacts of multi-scaling in wide area TCP traffic,” in *IEEE INFOCOM 00*, Tel Aviv, Israel, 2000.

[9] B. Flury, *A first course in Multivariate Statistics*. Springer Verlag, 1997.

[10] S. A. N. Ostring and H. Sirisena, “The influence of long-range dependence on traffic prediction,” in *Proceedings of ICC 2001*, 2001.

[11] D. Morato, J. Aracil, M. Izal, E. Magana, and L. A. Diez-Marca, “On linear prediction of Internet traffic for packet and burst switching networks,” in *Proceedings of IEEE International Conference on Computers, Communications and Networks*, October 2001.

[12] H. Saito, *Teletraffic technologies in ATM networks*. Artech-House, 1994.

[13] N. Johnson, *Continuous Univariate Distributions*. John Wiley and Sons, 1994, vol. 1.

[14] A. Sang and S. Li, “A predictability analysis of network traffic,” in *Proceedings of IEEE INFOCOM 2000*, 2000.