Blocking Behavior of Queueing System with Finite Buffer and Pseudo Self-similar Input

Shoji Kasahara
Graduate School of Information Science, Nara Institute of Science and Technology
Takayama 8916-5, Ikoma, Nara 630-0101, Japan
E-mail: kasahara@is.aist-nara.ac.jp

Abstract—*It is well known that the fitting methods based on the second-order statistics of counts for the arrival process are not sufficient for predicting the performance of the queueing system with self-similar input. However recent studies have revealed that the loss probability of finite queuing system can be well approximated by the Markovian input models. This paper studies the time-scale impact on the loss probability of $MMPP/D/1/K$ system where the MMPP is generated so as to match the variance of the self-similar process over specified time-scales. We investigate the loss probability changing system size, Hurst parameters and time-scales. Numerical examples show that the loss probability of $MMPP/D/1/K$ are not significantly affected by the time-scale.*

1 Introduction

The long-range dependency (LRD) and the self-similar process have been greatly concerned with the high speed computer communication networks [2, 4]. The self-similar process is quite different from traditional traffic models such as Poisson process and Markovian arrival process (*MAP*). This observation has initiated studies of new models such as chaotic maps, fractional Brownian motion (FBM) and fractional autoregressive integrated moving average (FARIMA) [2]. They can describe the self-similar behavior in a relatively simple manner. However, queueing theoretical techniques developed in the past are hardly applicable for these models.

On the other hand, a number of models based on traditional traffic models have been proposed. One approach is to emulate self-similarity over a certain range of time-scales with finite state Markovian models. [1] has proposed the model based on Markov-modulated Poisson process (*MMPP*) as a superposition of two state Markov processes. In [6], we have proposed a fitting method for self-similar traffic in terms of *MMPP*. Our fitting method is based on [1] where traffic is modeled by the superposition of several two state *MMPP*s. [1] proposed the fitting method which is mainly focused on the covariance structure of the second-order self-similar process. More precisely, the parameters of *MMPP* are determined so as to match the autocorrelation function which is approximately evaluated. In our method, however, the parameters are determined so as to match the variance of the self-similar process which is exactly evaluated.

It is well known that the queueing performance deteriorates with the self-similar traffic whose Hurst parameter is between 0.5 and 1, i.e., LRD traffic [2]. Therefore it is important to predict the queueing behavior under the self-similar traffic with LRD. In general, the fitting methods based on the second-order statistics of counts for the arrival process are not sufficient for predicting the queueing performance. In our previous study [6], numerical results showed that our fitting method works well in the sense of imitating statistical characteristics of self-similar traffic and that it does not work well for predicting the queueing behavior with resulting $MMPP/D/1$ even when the Hurst parameter is moderate. However, our fitting method succeeded in giving the tight upper bounds of the tail distribution of queue length.

Recently, [3] discussed the impact of the LRD on the buffer occupancy and indicated that LRD does not affect the buffer occupancy when the busy periods of the system are not large. In [5], the authors considered the critical time scale (CTS) and showed that the buffer behavior at the time-scale beyond the CTS is not significantly affected. Therefore, there is a possibility to predict the queueing behavior of system with LRD input using the fitting method based on the second-order statistics of counts for the arrival process if the appropriate time-scale is taken into consideration.

In this paper, we consider the time-scale impact on the queue length distribution using an *MMPP* generated so as to match the variance over specified time-scales. We use our fitting method for imitating the self-similar traffic with LRD and consider the resulting $MMPP/D/1/K$ system. We investigate the loss probability changing Hurst parameter, system size and time-scales.

This paper is organized as follows. In Section 2, we present the variance fitting method for the exact self-similar process. In Section 3, we summarize the analytical results of $MAP/G/1/K$, which is the generalization of $MMPP/D/1/K$. In Section 4, we show some numerical examples for the loss probability of $MMPP/D/1/K$ and illustrate the effects of time-scale, Hurst parameter and the system size. Finally, some concluding remarks are given in Section 5.

2 Variance Fitting Method

In this section we describe the process of determining the parameters of *MMPP*s so as to mimic the self-similar process. The readers are referred to [6] for details.

We consider a second-order stationary process $X = \{X_t : t = 1, 2, ...\}$ with the variance σ^2. In the context of the packet traffic, X_t corresponds to the number of packets that arrive during the t-th time slot. We also consider a new sequence of $X_t^{(m)}$ which is obtained by averaging the original sequence in non-overlapping blocks

$$X_t^{(m)} = \frac{1}{m}\sum_{i=1}^{m} X_{(t-1)m+i}, \quad t = 1, 2, \ldots .$$

We can define the self-similar process in terms of the variance of the averaged process. The readers are referred to [6, 9] for details.

Definition 2.1 *X is called exactly second-order self-similar with the Hurst parameter $H = 1 - \beta/2$ if*

$$Var(X^{(m)}) = \sigma^2 m^{-\beta}. \tag{1}$$

We construct an *MMPP* with self-similar behavior over several time-scales by superposing several *MMPP*s. First, consider two-state *MMPP*s with different time-scales. That is, the mean sojourn time of each process is in accordance with the different time-scale. Let us superpose them to make a new *MMPP*. When we see this process in a large time-scale, it looks like an ordinary two-state *MMPP*. If we look in a smaller time-scale, we find that each state is composed of a smaller *MMPP*. This can be repeated only a finite number of times.

The resulting *MMPP* is not self-similar from Definition 2.1 since it looks constant when the time-scale is larger than the time constant taken into consideration. However it can emulate self-similarity over several time-scales. Thus, it is practically sufficient to use the process which has self-similarity over only several time-scales to model the real traffic.

We assume that the number of states of each underlying *MMPP* is two. Therefore the *MMPP* obtained by the above method is also described by the superposition of several interrupted Poisson processes (*IPP*) and one Poisson process. We assume that the *MMPP* under consideration consists of $d(> 1)$ *IPP*s and a Poisson process. We also assume that two modulating parameters of each *IPP* are equal. Then we can describe ith *IPP* as follows

$$\boldsymbol{Q}_i = \begin{bmatrix} -\sigma_i & \sigma_i \\ \sigma_i & -\sigma_i \end{bmatrix}, \quad \boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_i & 0 \\ 0 & 0 \end{bmatrix}, \quad (1 \le i \le d).$$

Hence the superposition can be described as follows

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{Q}_1 \bigoplus \boldsymbol{Q}_2 \bigoplus \cdots \bigoplus \boldsymbol{Q}_d, \\ \boldsymbol{\Lambda} &= \boldsymbol{\Lambda}_1 \bigoplus \boldsymbol{\Lambda}_2 \bigoplus \cdots \bigoplus \boldsymbol{\Lambda}_d \bigoplus \lambda_p, \end{aligned}$$

where $\bigoplus$ means the Kronecker's sum and λ_p is the arrival rate of the Poisson process to be superposed. The whole arrival rate of the process λ is given by

$$\lambda = \lambda_p + \sum_{i=1}^{d} \frac{\lambda_i}{2}. \tag{2}$$

Parameters which we have to determine are σ_i, $\lambda_i (1 \le i \le d)$, and λ_p.

First, as preliminary we define the notations used in the procedure and describe some assumptions. Let $N_{t|i}$ be the number of arrivals in the i-th *IPP* during the tth time slot and $N_{t|p}$ be the number of arrivals in the Poisson process, and let $N_{t|i}^{(m)}$ and $N_{t|p}^{(m)}$ be respectively the averaged processes of $N_{t|i}$ and $N_{t|p}$. We assume that

$$\mathrm{Var}\left(X_t^{(m)}\right) = \mathrm{Var}\left(\sum_{i=1}^{d} N_{t|i}^{(m)} + N_{t|p}^{(m)}\right).$$

We obtain the variance of the i-th *IPP* as [6]

$$\begin{aligned} \mathrm{Var}\left(N_{t|i}^{(m)}\right) &= \frac{\lambda_i}{2m} \\ &+ \left\{\frac{1}{4m\sigma_i} - \frac{1}{8m^2\sigma_i^2}(1 - e^{-2m\sigma_i})\right\}\lambda_i^2. \end{aligned}$$

The corresponding variance of the Poisson process is λ_p/m. Because the variance of a process which is a superposition of independent subprocesses equals the sum of individual variances, the variance of the whole process is given by

$$\mathrm{Var}\left(X_t^{(m)}\right) = \frac{\lambda}{m} + \frac{1}{4}\sum_{i=1}^{d} \eta_i \lambda_i^2, \tag{3}$$

where

$$\eta_i = \frac{1}{m\sigma_i} - \frac{1}{2m^2\sigma_i^2}(1 - e^{-2m\sigma_i}).$$

Using (1) and (3), we match the variance at d different points m_i $(1 \le i \le d)$. Suppose the range of time-scales over which we want the process to express self-similarity of the original process is $m_{\min} \le m \le m_{\max}$, then m_i is defined by

$$m_i = m_{\min} a^{i-1} \quad (1 \le i \le d),$$

where

$$a = (m_{\max}/m_{\min})^{1/(d-1)}, \qquad d > 1. \tag{4}$$

From the property of η_i, we have

$$\lambda/m < \mathrm{Var}\left(X_t^{(m)}\right) < \lambda/m + \lambda^2. \tag{5}$$

We must choose m_i such that (5) is satisfied at any m_i. This condition comes from that we use a simple *IPP* as a sub-process. Practically, this condition never matters when m is large, but sometimes $\mathrm{Var}(X_t^{(m)})$ is too small when m is small. Therefore, we should be careful to choose m_1, which is large enough to make $\mathrm{Var}(X_t^{(m_1)})$ larger than $\lambda/m_{\min}$.

Next, we assume that σ_i can be described as

$$\sigma_i = \sigma_1 m_1 / m_i \quad (1 \le i \le d). \tag{6}$$

This assumption comes from the intuitive understanding that a self-similar process looks the same in any time-scale. By this assumption, we can reduce the number of the parameters to be determined. That is, if we determine σ_1, we can obtain the values of σ_i $(2 \le i \le d)$ by using (6). Furthermore, we can obtain λ_p from the following equation if we determine λ_i.

$$\lambda = \lambda_p + \sum_{i=1}^{d} \frac{\lambda_i}{2}.$$

Now the parameters we need to find are only σ_1 and λ_i.

In the following, we describe the procedure of determining these parameters. We show the parameters preliminarily required for our fitting procedure in Table 1.

Procedure of Parameter Fitting

Step 0. Find the range of σ_1 heuristically and fix σ_1.

Step 1. Determine λ_i as the function of σ_i. From (1) and (3), we have

$$\sigma^2 \begin{bmatrix} m_1^{-\beta} \\ m_2^{-\beta} \\ \vdots \\ m_d^{-\beta} \end{bmatrix} = \lambda \begin{bmatrix} m_1^{-1} \\ m_2^{-1} \\ \vdots \\ m_d^{-1} \end{bmatrix} + \boldsymbol{B} \begin{bmatrix} \lambda_1^2 \\ \lambda_2^2 \\ \vdots \\ \lambda_d^2 \end{bmatrix}, \tag{7}$$

where $\boldsymbol{B}$ is the $d \times d$ matrix whose (i, j) element is

$$B_{ij} = \frac{1}{4m_i\sigma_j} - \frac{1}{8m_i^2\sigma_j^2}(1 - e^{-2m_i\sigma_j}). \tag{8}$$

Solving this, we determine λ_i as the function of σ_i.

Step 2. Determine the value of σ_1 from the range found in Step 0. Here we consider the integral of the difference between the log-scales variance curve of the process given by (3) and that of the self-similar process given by (1) over defined time-scales. We determine the value of σ_1 so as to minimize that integral.

Step 3. Determine the values of λ_i from (7).

In step 1, it is necessary that $\boldsymbol{B}$ is non-singular. It is difficult to prove the non-singularity of $\boldsymbol{B}$ for any positive integer of d, however, we can show that if a is sufficiently large, $\boldsymbol{B}$ is non-singular for any σ_1. We have discussed the non-singularity of the matrix $\boldsymbol{B}$ in [6].

When we minimize the integral in step 2, we must be careful to keep the values of λ_i and λ_p larger than zero. We consider the minimum at the log-scale because we can treat smaller time-scales more carefully.

Table 1: Preliminarily Required Parameters

Parameter	Meaning
λ	Arrival rate of the whole process
$m_{\min}$, $m_{\max}$	Minimum and maximum of time-scales over which self-similarity is taken into consideration
σ^2	Variance
H	Hurst parameter
d	Number of *IPP*s

3 *MAP/G/1/K*

In this section, we briefly summarize the analytical results of *MAP/G/1/K*, a single server queue with finite capacity and Markovian arrival process [7]. Note that *MMPP/D/1/K* is the special case of *MAP/G/1/K*.

In *MAP/G/1/K*, the customer arrives at the system according to the *MAP* represented by $(\boldsymbol{C}, \boldsymbol{D})$, where $\boldsymbol{C}$ and $\boldsymbol{D}$ are $M \times M$ matrices. The service time S is generally and identically distributed with the distribution function $S(x)$. The irreducible matrix $\boldsymbol{C} + \boldsymbol{D}$ is the infinitesimal generator of the underlying Markov process restricted to the states $\{1, \cdots, M\}$. Let $\boldsymbol{\pi}$ denote the stationary vector of $\boldsymbol{C} + \boldsymbol{D}$, i.e.

$$\boldsymbol{\pi}(\boldsymbol{C} + \boldsymbol{D}) = \boldsymbol{0}, \quad \boldsymbol{\pi e} = 1, \tag{9}$$

where $\boldsymbol{e}$ denotes the column vector of ones.

Let $\boldsymbol{A}_k$ $(k \geq 0)$ denote an $M \times M$ matrix whose (i, j)th element represents the conditional probability that k customers arrive at the system during a service time of a customer and the underlying Markov chain is in phase j at the end of the service given that the underlying Markov chain is in phase i at the beginning of the service. Then $\boldsymbol{A}_k$ satisfies the following equation

$$\sum_{k=0}^{\infty} \boldsymbol{A}_k z^k = \int_0^{\infty} e^{(\boldsymbol{C}+z\boldsymbol{D})x} dS(x).$$

We also define $\boldsymbol{B}_k$ $(k \geq 1)$ as

$$\boldsymbol{B}_k = \sum_{n=k}^{\infty} \boldsymbol{A}_k.$$

Let $\boldsymbol{x}_k$ $(0 \leq k \leq K-1)$ denote a $1 \times M$ vector whose ith element represents the stationary joint probability that the number of customers in the system at departures is k and the phase of the arrival process is i. $\boldsymbol{x}_k$ satisfies the following equations

$$\begin{aligned} \boldsymbol{x}_k &= \boldsymbol{x}_0 \boldsymbol{E} \boldsymbol{A}_k + \sum_{n=1}^{k+1} \boldsymbol{x}_n \boldsymbol{A}_{k+1-n}, \ 0 \leq k \leq K-2, \\ \boldsymbol{x}_{K-1} &= \boldsymbol{x}_0 \boldsymbol{E} \boldsymbol{B}_{K-1} + \sum_{n=1}^{K-1} \boldsymbol{x}_n \boldsymbol{B}_{K-n}, \end{aligned}$$

where $\boldsymbol{E} = (-\boldsymbol{C})^{-1}\boldsymbol{D}$. We can calculate $\boldsymbol{x}_k$'s in the iterative manner [7].

Let $\boldsymbol{y}_k$ $(0 \leq k \leq K)$ denote a $1 \times M$ vector whose ith element is the stationary joint probability that the number of customers in the system is k and the phase of the arrival process is i at an arbitrary time. We define the vector probability generating function $\boldsymbol{y}(z)$ as

$$\boldsymbol{y}(z) = \sum_{k=0}^{K} \boldsymbol{y}_k.$$

In general, the following relation between the probability generating function for arbitrary points and that for departure points is hold under any service disciplines [8]

$$\boldsymbol{y}(z)(\boldsymbol{C} + z\boldsymbol{D}) = \lambda(z-1)\boldsymbol{x}(z), \tag{10}$$

where λ is the mean arrival rate of the *MAP* and given by $\lambda = \boldsymbol{\pi D e}$. Suppose that all arriving customers can accommodate the system and that the customer who finds the system full immediately leave the system. Then $\boldsymbol{x}(z)$ in (10) is given by

$$\begin{aligned} \boldsymbol{x}(z) &= \frac{1 - \boldsymbol{y}_0\boldsymbol{e}}{\rho} \sum_{k=0}^{K-1} \boldsymbol{x}_k z^k \\ &\quad + \frac{\rho - (1 - \boldsymbol{y}_0\boldsymbol{e})}{\rho} \frac{\boldsymbol{y}_K \boldsymbol{D}}{\boldsymbol{y}_K \boldsymbol{D e}} z^K, \end{aligned} \tag{11}$$

where $\rho = \lambda E[S]$. Comparing the coefficients of z^0 in (11), we obtain

$$\boldsymbol{y}_0 = \lambda \frac{1 - \boldsymbol{y}_0 \boldsymbol{e}}{\rho} \boldsymbol{x}_0 (-\boldsymbol{C})^{-1}.$$

Multiplying both sides of above equation by $\boldsymbol{e}$ yields

$$\boldsymbol{y}_0 \boldsymbol{e} = \frac{\boldsymbol{x}_0 (-\boldsymbol{C})^{-1}}{E[S] + \boldsymbol{x}_0 (-\boldsymbol{C})^{-1}}.$$

Similarly, comparing the coefficients of z^k ($k = 1, \cdots, K-1$) in (11) yields

$$\boldsymbol{y}_k \boldsymbol{C} + \boldsymbol{y}_{k-1} \boldsymbol{D} = \frac{1 - \boldsymbol{y}_0 \boldsymbol{e}}{E[S]} (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k).$$

From this, $\boldsymbol{y}_k$ ($k = 1, \cdots, K-1$) can be calculated by

$$\boldsymbol{y}_k = \boldsymbol{y}_{k-1} \boldsymbol{D} (-\boldsymbol{C})^{-1} + \frac{1 - \boldsymbol{y}_0 \boldsymbol{e}}{E[S]} (\boldsymbol{x}_{k-1} - \boldsymbol{x}_k)(-\boldsymbol{C})^{-1}.$$

Finally, we obtain $\boldsymbol{y}_K$ from

$$\boldsymbol{y}_K = \pi - \sum_{k=0}^{K-1} \boldsymbol{y}_k.$$

The loss probability P_l is defined as the ratio of lost customers to the arriving customers. Then P_l satisfies the following relationship between the server utilization and the offered load

$$1 - \boldsymbol{y}_0 \boldsymbol{e} = \rho (1 - P_l).$$

Therefore P_l is given by

$$P_l = 1 - (1 - \boldsymbol{y}_0 \boldsymbol{e})/\rho. \tag{12}$$

4 Numerical Results

In this section, we show some numerical examples of the loss probability for $MMPP/D/1K$ under several conditions.

We calculate P_l according to the way of Section 3 and [7]. Here we set the followings:

$$\boldsymbol{C} = \boldsymbol{Q} - \boldsymbol{\Lambda}, \quad \boldsymbol{D} = \boldsymbol{\Lambda},$$

where $\boldsymbol{Q}$ is the infinitesimal generator of $MMPP$ and $\boldsymbol{\Lambda}$ is the arrival rate matrix. $\boldsymbol{Q}$ and $\boldsymbol{\Lambda}$ are calculated by our fitting method described in Section 2. Table 2 shows the values of preliminary parameters for our fitting method. We calculate P_l changing the mean service time, i.e., offered load ρ. We also investigate the behavior of P_l changing the system size K. In all figures, TS denotes the common logarithm of the time-scale, K the system size, and ρ the offered load.

Fig. 1 represents the loss probability with $TS = 5$ and $K = 30$. The loss probability is calculated with changing H and ρ. We observe that the loss probability becomes large as H and ρ are getting large.

Fig. 2 shows the loss probability with $K = 50$ and $H = 0.6$. The loss probability is calculated with changing TS and ρ. From Fig. 2, we observe that the loss probability becomes large as ρ becomes large. However there is little difference of the loss probabilities in terms of the time-scale.

Table 2: Parameters

Parameter	Value
λ	1.0
σ^2	0.6
$m_{\min}$	10^2
$m_{\max}$	10^4, 10^5, 10^6, 10^7
H	0.6, 0.7, 0.8, 0.9
d	4

Fig. 3 illustrates the loss probability with $\rho = 1.0$ and $K = 100$. The loss probability is calculated with changing TS and H. From the figure, we observe that the loss probability increases gradually when H becomes large. We also observe that the system size affects the loss probability. However, there are no great differences when the time-scale changes.

Fig. 4 represents the loss probability with $\rho = 1.0$ and $H = 0.8$, we observe that the loss probability decreases gradually when K becomes large. However it does not change greatly by the time-scale.

Fig. 5 illustrates the loss probability with $TS = 4$ and $K = 30$. The loss probability is calculated with changing H and ρ. We observe that the loss probability becomes large as H and ρ are getting large. This tendency is similar to Fig. 1. That is, the loss probability is affected by H and ρ.

5 Conclusion

In this paper, we investigated the loss probability of $MMPP/D/1/K$ where $MMPP$ is generated so as to mimic the variance-time curve of the self-similar process over several time-scales. From the numerical examples, we observed that the loss probability is sensitive to the offered load, system size and Hurst parameter while not sensitive to the time-scale.

Our numerical results represent a significant meaning when we consider the modeling and analysis of the queueing system with self-similar input. As we stated in Introduction, it is known that the fitting methods based on the second-order statistics of counts for the arrival process are not sufficient for predicting the queueing performance correctly. However, this is the case of queueing system with infinite buffer and not the case of that with finite buffer. Our numerical results reveal that the time-scale does not have a strong impact on the loss probability for a certain parameter values of system size, Hurst parameter and traffic intensity. Therefore, it seems to be possible to predict the blocking behavior of the finite queuing system with LRD inputs by the pseudo self-similar process.

Currently we are performing the further numerical studies for finding the critical values of the system size and the time-scales in terms of the loss probability.

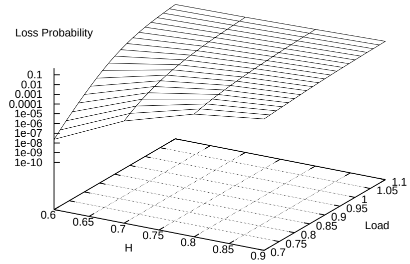

Figure 1: Loss Probability ($K = 30$, $TS = 5$)

References

[1] A. T. Anderson and B. F. Nielsen, "A Markovian Approach for Modeling Packet Traffic with Long-Range Dependence," *IEEE J. Select. Areas Commun*, vol. 16, no. 5, pp.719-732, June 1998.

[2] A. Erramilli, O. Narayan and W. Willinger, "Fractal Queueing Models," *Frontiers in Queueing*, ed. J. H. Dshalalow, CRC Press, Inc., pp.245-269, 1997.

[3] D. P. Heyman and T. V. Lakshman, "What Are the Implications of Long-Range Dependence for VBR-Video Traffic Engineering?," *IEEE/ACM Trans. Networking*, vol. 4, no. 3, pp.301-317, June 1996.

[4] W. Leland, M. Taqqu, W. Willinger and D. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp.1-15, 1994.

[5] B. K. Ryu and A. Elwalid, "The Importance of Long-Range Dependence of VBR Video Traffic in ATM Traffic Engineering: Myths and Realities," *ACM SIGCOMM'96*, pp.3-14, 1996.

[6] T. Yoshihara, S. Kasahara, and Y. Takahashi, "Practical Time-scale Fitting of Self-similar Traffic with Markov-modulated Poisson Process," To appear in *Telecommunication Systems*, 2000.

[7] T. Takine, "Introduction to Numerical Solution Techniques of Finite-State Markov Chains and Its Applications," Preprint, 1995 (in Japanese).

[8] T. Takine and Y. Takahashi, "On the Relationship between Queue Lengths at a Random Instant and at a Departure in the Stationary Queue with BMAP Arrivals," Stochastic Models, vol.14, no.3, pp.601-610, 1998.

[9] B. Tsybakov and N. D. Georganas, "On Self-Similar Traffic in ATM Queues: Definitions, Overflow Probability Bound, and Cell Delay Distribution," *IEEE/ACM Trans. Networking*, vol. 5, no. 3, pp.397-409, 1997.

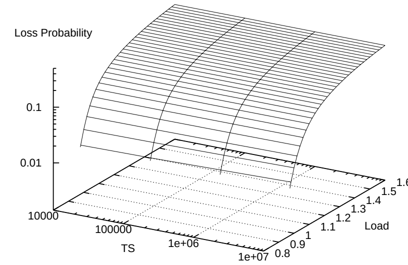

Figure 2: Loss Probability ($H = 0.6$, $K = 50$)

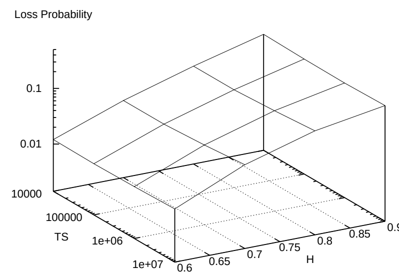

Figure 3: Loss Probability ($\rho = 1.0$, $K = 100$)

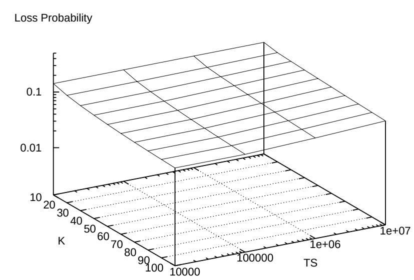

Figure 4: Loss Probability ($\rho = 1.0$, $H = 0.8$)

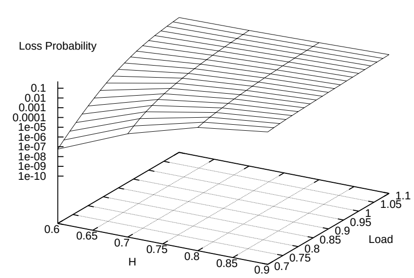

Figure 5: Loss Probability ($K = 30$, $TS = 4$)