*Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.*

# BLOCKING IN TANDEM QUEUES

**Brigitte NICOLAS and Guy LATOUCHE**

**Université Libre de Bruxelles**
**Séminaire de Théorie des Probabilités, CP 212**
**Boulevard du Triomphe**
**1050 Bruxelles**
**Belgium**

**ABSTRACT**

We consider a system of two queues in tandem with a finite intermediary buffer. We examine the influence of variability in service requirement at the second server, on the behaviour of the system.

*Introduction*

The queueing model considered here consists of two units in series with a finite intermediary buffer. Arriving customers enter an infinite buffer in front of Unit I. After being served at Unit I, a customer enters a finite buffer in front of Unit II if the buffer is not full; if the buffer is full, the customer under consideration may not leave Unit I, which thereby becomes blocked and unable to process waiting customer. At a later time, Unit I becomes available again, in a manner to be described later.

In many practical situations, e.g. in data communication networks, the use of an intermediary buffer is dictated by the physical necessity of decoupling the functioning of Units I and II. In other circumstances, it may be advantageous to use an intermediary buffer, in order to render each unit less dependent on random fluctuations in the functioning of the other.

Our purpose in the present note is to examine how the variability of service requirements at Unit II influences the functioning of Unit I. The mathematical model and method of analysis are described in the next section. In Section 2 are defined the parameter values chosen for the numerical analysis. Some results are presented and discussed in Section 3.

For a survey of the literature on such systems, we refer to the bibliography in Latouche and Neuts [1], where a similar system is studied.

1. *The mathematical model*

We assume that customers arrive in the system according to a Poisson process with parameter $\lambda$; the duration of service at Unit I is exponential with parameter $\mu$; the service at Unit II is phase-type (PH) with representation $(\underline{\alpha},T)$; all random variables are independent.

PH distributions form a general class, defined and extensively analysed in Neuts [2]. In short, a random variable has a PH distribution if it may be represented as the time until absorption for a Markov process with one absorbing state. They are characterized by the number m of transient states

(or phases), the stochastic m-vector $\underline{\alpha}$ that gives the initial probability distribution on the transient states, and the infinitesimal generator T, of order m, that determines transitions among the transient states. Erlang, hyperexponential and Coxian distributions with real, positive parameters, all are special cases of PH distributions.

Customers who have not yet been served at Unit I are called 1-customers; customers who have been served at Unit I but not at Unit II are called 2-customers. The intermediary buffer is finite and we denote by M the maximum number of 2-customers : there are at most M-2 such customers in the buffer, one being served at Unit II, and one unable to leave Unit I when the buffer is full.

We assume that when Unit I becomes blocked, it stays so until there remain K 2-customers in the system, $0 \leqslant K \leqslant M-1$. The case when K=M-1 means that Unit I operates again as soon as one 2-customer finishes its service, thereby releasing one space in the buffer and allowing the blocking customer to leave Unit I. The case when K=0 means that the buffer and Unit II must become completely empty before Unit I may start functioning again.

The quantity M-2 may be thought of as a technological constraint on the buffer size, while K determines a control policy, to be used e.g. if there are costs associated to the shutting off and starting up of Unit I.

Under the stated assumptions, the system may be described as a Markov process on the state space $\{(n,i,j); n \geqslant 0,\ i=0,1,\ldots,M-1,M',(M-1)',\ldots,(K+1)';\ 1 \leqslant j \leqslant m\}$, where n is the number of 1-customers, i is indicative of the number of 2-customers and the state of Unit I (a symbol ' meaning that Unit I is blocked), and j is the service phase at Unit II. Since n may change by one unit at most, that Markov process is a quasy-birth-and death process of the type extensively studied by Neuts [2]. The corresponding analysis is well documented in the literature and shall not be reproduced here, for lack of space. The interested reader will find the general theorems in [2].
In Nicolas [3], the theory has been applied to the model at hand, and several specific results have been obtained. To obtain the stationary probability distribution, it is necessary to compute a matrix of order N=(2M-K); the algorithm developed in [3] is such that no other matrix of that order need be stored.

## 2. *The numerical analysis*

The stochastic process is specified by the following parameters :

- the input rate $\lambda$,
- the service rate $\mu$ at Unit I;
- the maximum number M of 2-customers;
- the control parameter K;
- the representation $(\underline{\alpha},T)$ of the service distribution at Unit II.

Our purpose is to measure how variability in service requirements at Unit II influences the bahaviour of the system. A frequently used, global measure of variability is the ratio C of the standard deviation to the expected value. We have constructed, for a number of values of C, several PH-distributions with the same value for C. For each, the expected value equals one, thereby the unit of time is fixed.

The queueing system is stable if and only if $\lambda < \lambda_{max}$, where $\lambda_{max}$ is a non explicit, but easily computed, function of all the other parameters. This we have firstly examined.

We then have studied, for certain values of $\lambda$, the queue in front of Unit I and Unit II and both the stationary probability $\pi_o$ that Unit I is blocked and the stationary probability $\pi_1$ that Unit I becomes blocked at the end of a service : $\pi_o$ and $\pi_1$ give different information since the former is a time-average, while the later is a customer-average.

## 3. *Numerical results*

*3.1 The maximal arrival rate.* We have systematically observed that it is an increasing function of K, for fixed $\mu$, M and $(\underline{\alpha},T)$, therefore it is best to set K=M-1 in order to maximize the throughput of the system. Also, it is a monotonically increasing function of $\mu$, for fixed K, M and $(\underline{\alpha},T)$ : see Figures 1 and 2 where different PH-distributions are identified by their coefficient of variability C. Not surprising, $\lambda_{max}$ is bounded aboved by the minimum of $\mu$ and $(E\ [\text{service at Unit II}])^{-1}$.

3.2 *The stationary distribution of the system* is similarly affected by the variability of the PH-distribution $(\underline{\alpha},T)$. We display on Figures 3,4 and 5 respectively values of $m_1$, $m_2$ and $\pi_o$, where $m_i$=E (number of i-customers), i=1,2, and $\pi_o$ is the stationary probability that Unit I is blocked.

To compare different systems under "equal load" conditions, one may either impose the same rate of arrival $\lambda$, or the same ratio $\rho=\lambda/\lambda_{max}$. Because of the large differences in $\lambda_{max}$ for different distributions $(\underline{\alpha},T)$, one may not confuse the two definitions. Since $\lambda$ is the real systems parameter, we display $m_1$, $m_2$ and $\pi_o$ as functions of $\lambda$. In order to keep part of the information that might be contained in $\rho$, we have marked each curve by dots corresponding to the values $\rho$=.3,.5,.7 and .9.

We observe on Figure 4 one instance when it is necessary to properly define the notion of equal load. For a given value of $\lambda$, $m_2$ clearly increases with the variability of the PH-distribution. For fixed $\rho$=0.9 however, the values of $m_2$ for each distribution (highest point on each curve) are nearly equal.

We must emphasize that the behaviour of the system depends on the whole distribution $(\underline{\alpha},T)$, and not on the coefficient C only. Results not reproduced here indicate the existence of distributions $(\underline{\alpha},T_1)$ and $(\underline{\alpha},T_2)$ such that $C_{(1)} > C_{(2)}$ but $m_{1(1)} < m_{1(2)}$, $m_{2(1)} < m_{2(2)}$ and $\pi_{o(1)} < \pi_{o(2)}$ for whole ranges of values of $\lambda$.

3.3 *The blocking phenomenon* may be measured either by the stationary probability $\pi_o$ that Unit I is blocked at time t, or the stationary probability $\pi_1$ that Unit I becomes blocked after serving a 1-customer. The former measures the length of time spent in the blocked state, the latter measures the frequency of switching from a state where Unit I is available to the state where it is blocked.

The variability in ($\underline{\alpha}$,T) influences differently $\pi_o$ and $\pi_1$ (see Figures 5 and 6). For instance, consider the values of $\pi_o$ and $\pi_1$, for C=2.5 and 5 respectively, and $\lambda$=5. We observe that the queue with highest variability is blocked during longer periods of time ($\pi_o(C=5) > \pi_o(C=2.5)$) but changes less frequently from being available to being blocked ($\pi_1(C=5) < \pi_1(C=2.5)$).

This may be interpreted as follows. For the distribution with C=5, services at Unit II are typically very short, with an occasional very long one. When a very long service occurs, Unit I is likely to become blocked and to remain so for a long time. When that long service terminates, Unit II will process many customers with a very short service, during which time Unit I is unlikely to become blocked again.

*REFERENCES*


[1] LATOUCHE, G. and NEUTS, M.F. : "Efficient algorithmic solutions to exponential tandem queues with blocking", SIAM J. Alg. Disc. Methods 1, 1980, 93-106.

[2] NEUTS, M.F. : "Matrix-Geometric Solutions in Stochastic Models. An algorithmic Approach", The Johns Hopkins University Press, Baltimore, 1981

[3] NICOLAS, B. : "Analyse quantitative d'un système de deux files en tandem avec blocage", Mémoire de licence en mathématiques, Université Libre de Bruxelles, 1981

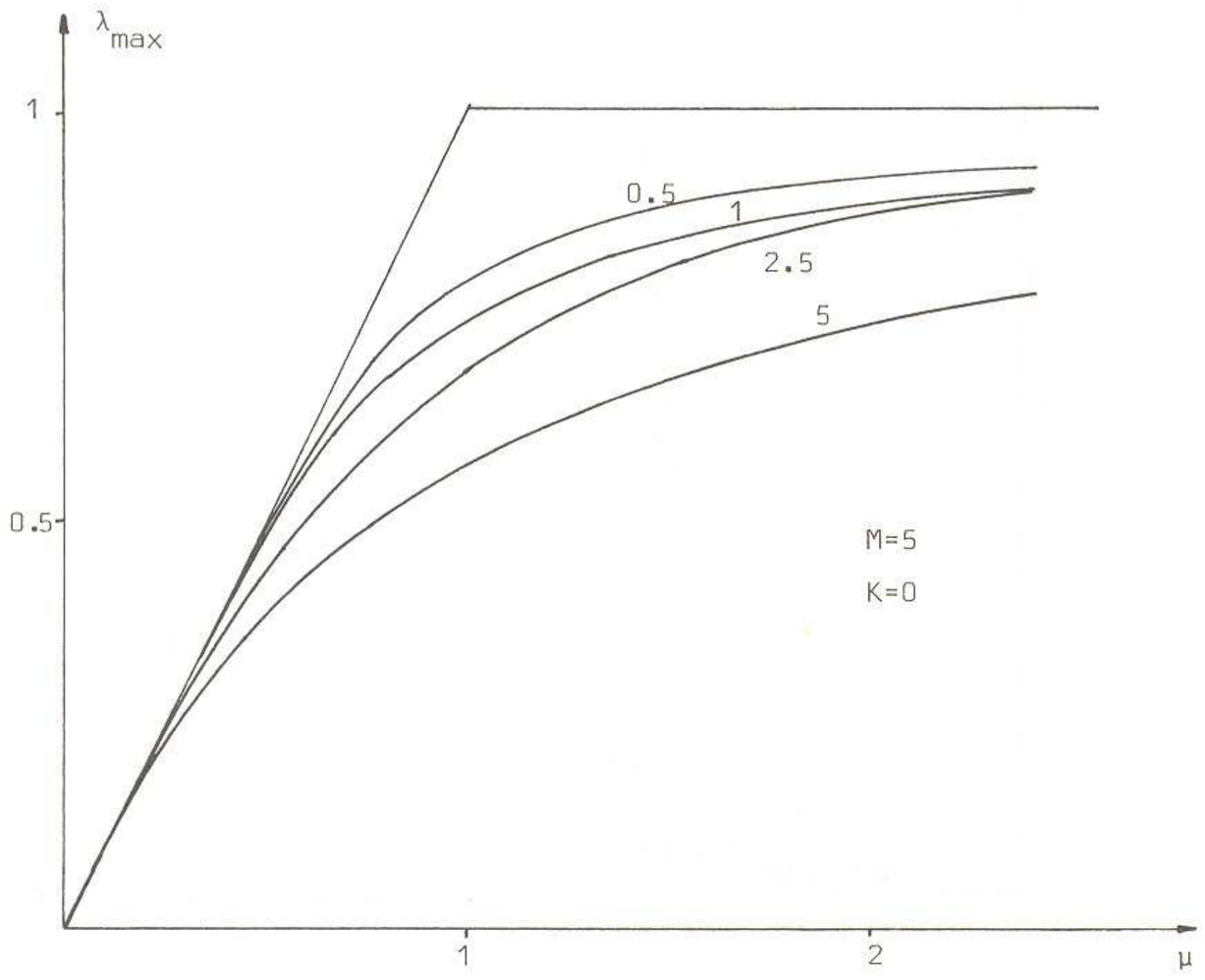


Figure 1. Maximum throughput, K=0

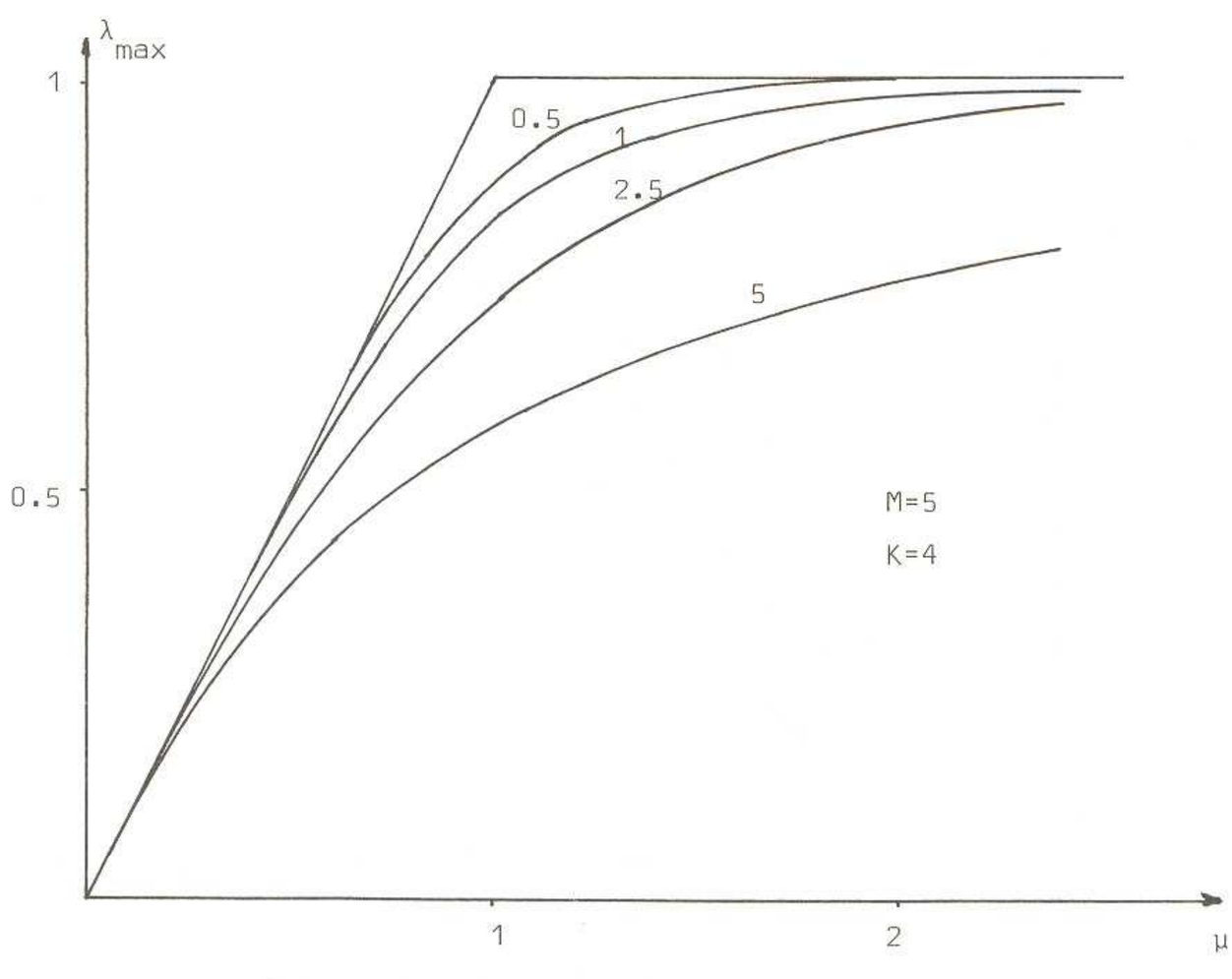


Figure 2. Maximum throughput, K=M-1

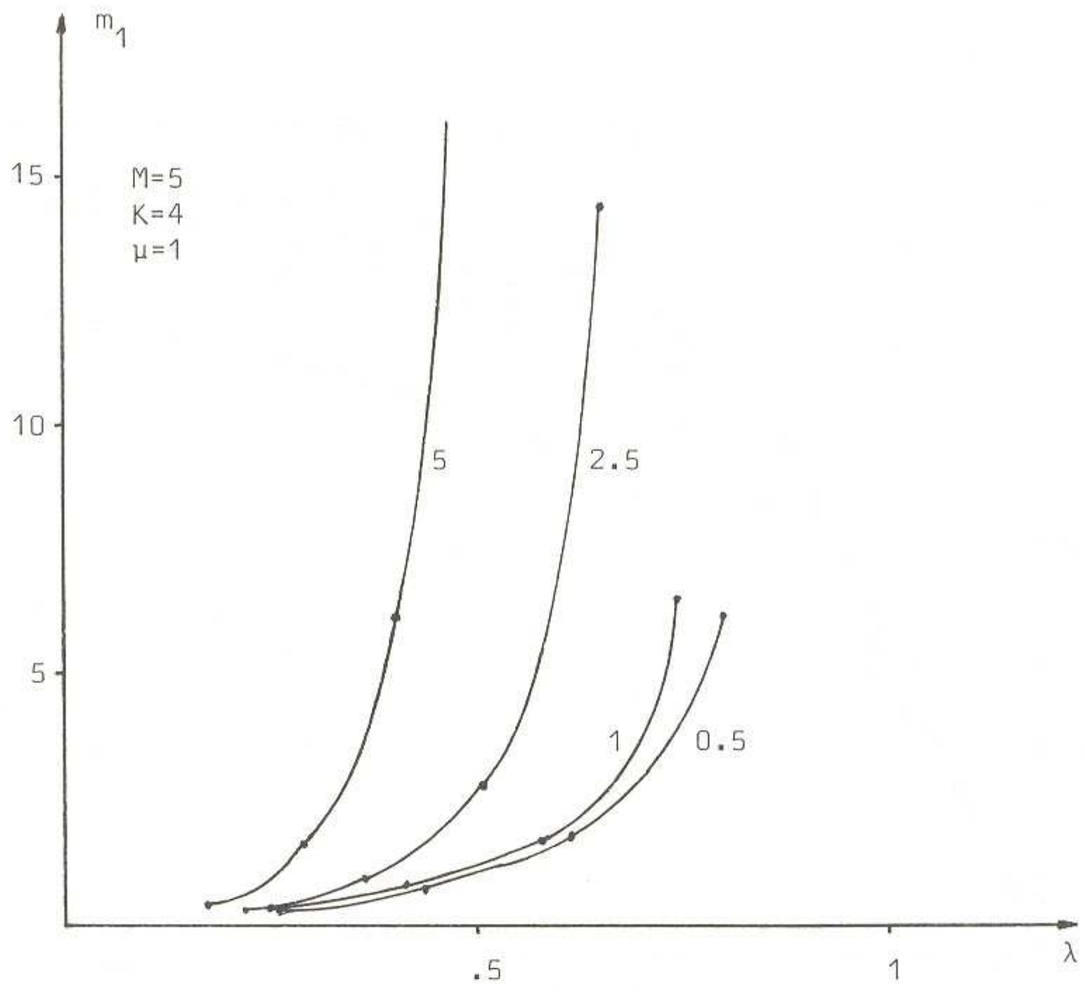


Figure 3. Expected number of 1-customers

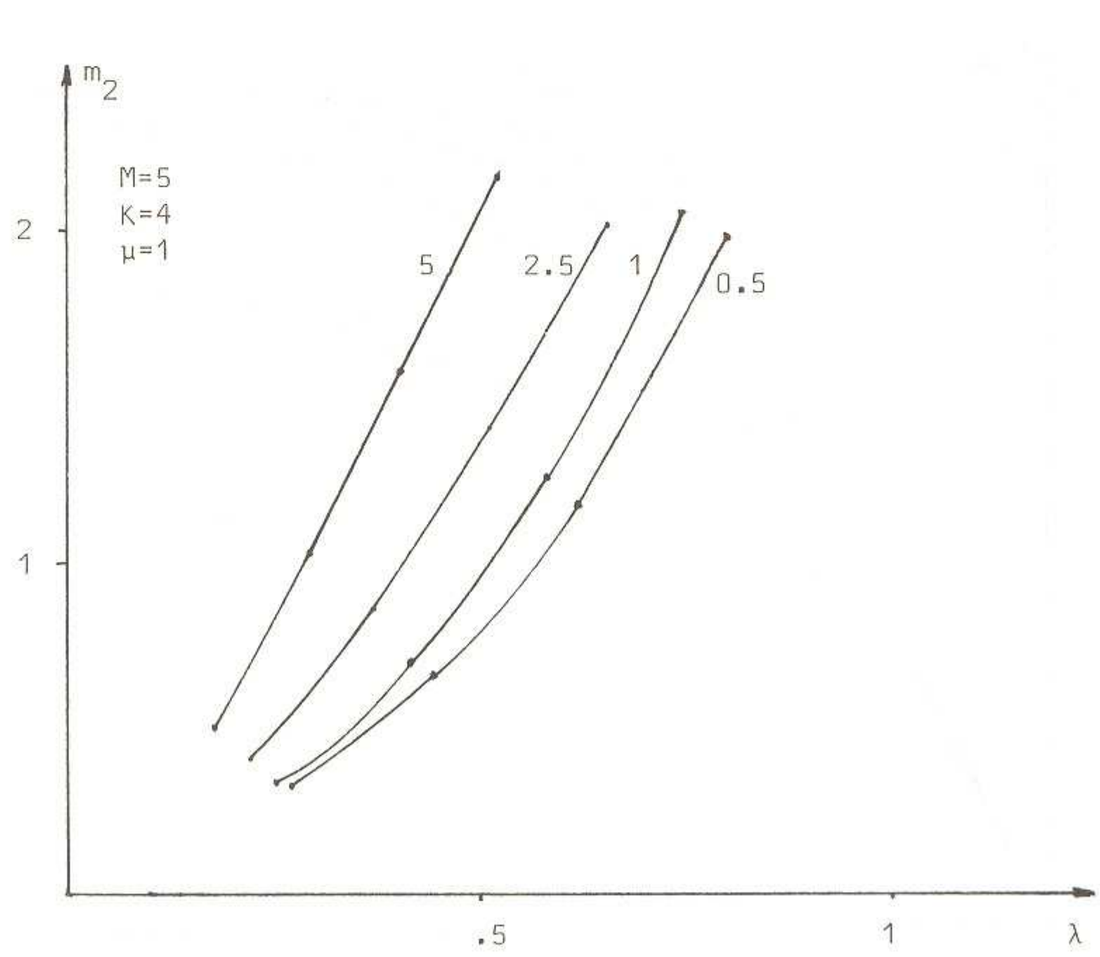


Figure 4. Expected number of 2-customers

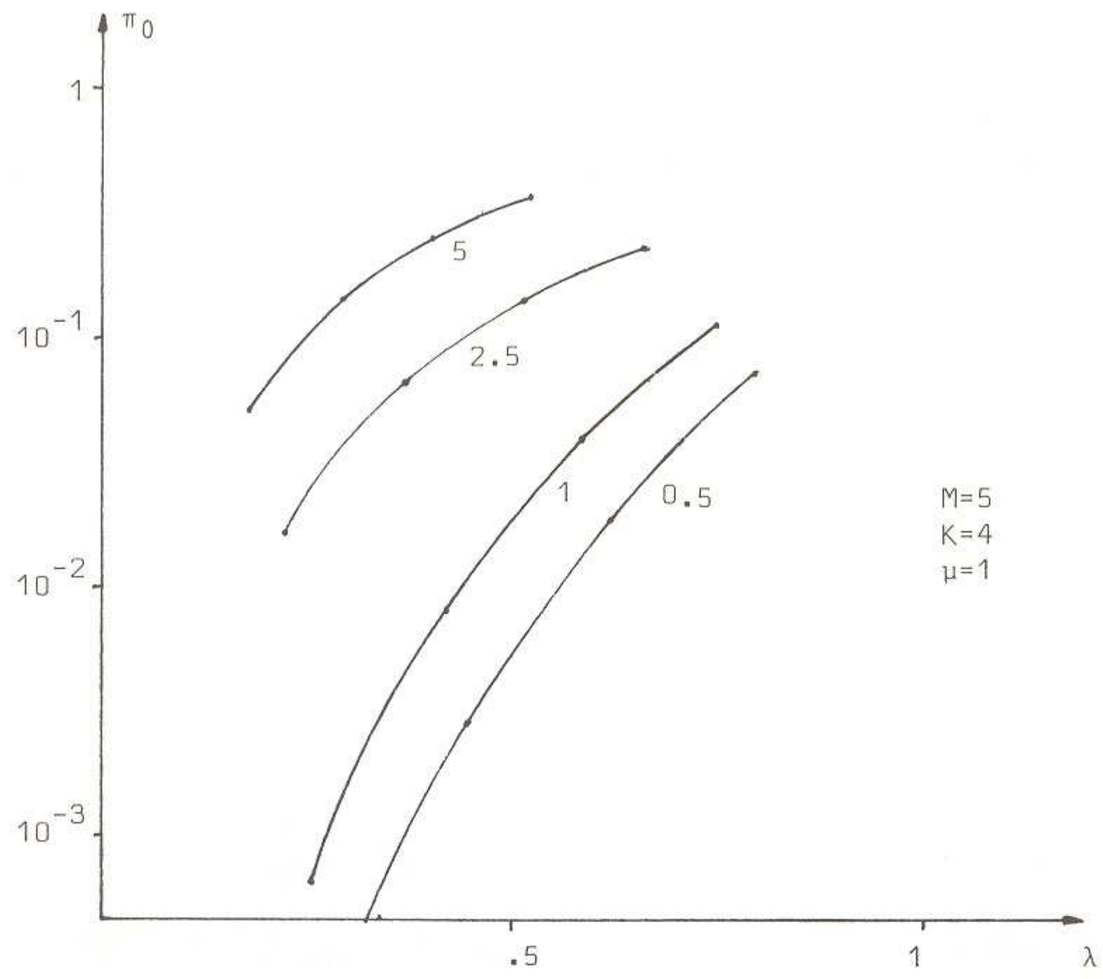


Figure 5. Probability that Unit I is blocked

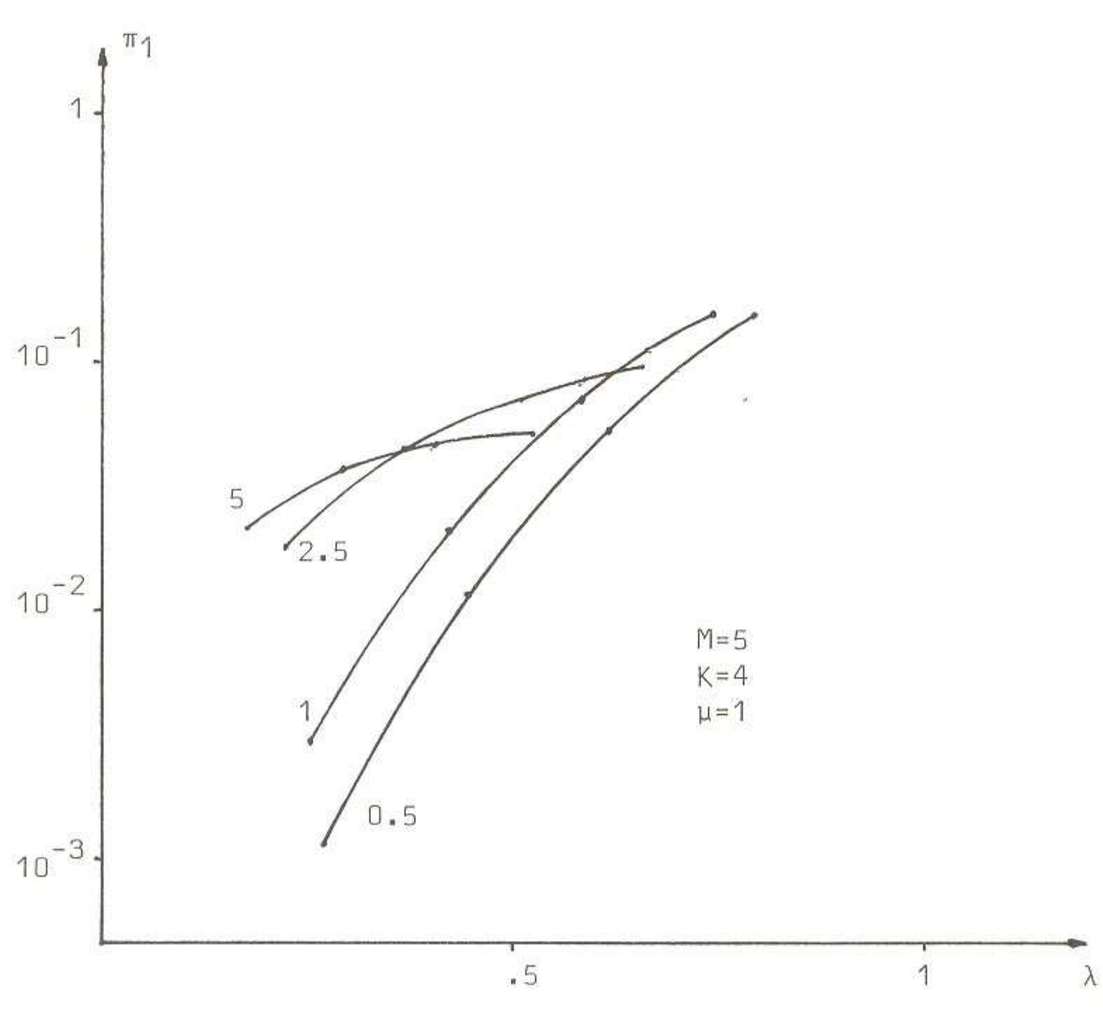


Figure 6. Probability that Unit I becomes blocked