PUC

Series: Monografias em Ciência da Computação
Nº 2/79

UPPER AND LOWER BOUNDS FOR THE WAITING TIME IN THE M/M/m AND M/G/m QUEUEING SYSTEMS

by

Gunter H. Bolch

Departamento de Informática

Pontifícia Universidade Católica do Rio de Janeiro
Rua Marquês de São Vicente, 225 - CEP 22453
Rio de Janeiro — Brasil

PUC/RJ - DEPARTAMENTO DE INFORMÁTICA

Series: Monografias em Ciência da Computação
Nº 2/79

Series Editor: Daniel A. Menascé February, 1979

UPPER AND LOWER BOUNDS FOR THE WAITING TIME IN M/M/m AND M/G/m QUEUEING SYSTEMS*

by

Gunter R.H. Bolch

* This work has been supported in part by FINEP.

Abstract

In the modelling of computer systems the waiting time in M/M/m and M/G/m queueing systems is a very important variable.

There exists a formula for the waiting time in M/M/m queueing systems. But this formula is very complex compared with the formula for a M/M/1 queueing system.

For M/G/m queueing systems there exists no formula for the waiting time. In this case one can only approximate the waiting time by using the M/M/m formula or the existing formulae for the upper and the lower bounds of the waiting time in G/G/m queueing systems.

In the first part of the paper formulae for an upper and a lower bound are developed and from those an approximate formula for the waiting time in M/M/m queueing systems is derived. These formulae are not more complicated than the M/M/1 waiting time formula.

These results are extended to the M/G/m queueing system in the second part. It is shown that these are much better than the results by an M/M/m approximation and that the new bounds are within the known bounds for G/G/m queueing systems.

Keywords

Multi server queueing system, waiting time, queue length, upper and lower bounds, approximation.

Resumo

O tempo de espera dos sistemas de fila M/M/m e M/G/m é uma variável muito importante na modelagem de sistemas de computação.

Existe uma fórmula para o tempo de espera no sistema M/M/m, porém esta fórmula é muito complexa comparada a do sistema M/M/1.

Para o sistema M/G/m nao existe fórmula alguma para o tempo de espera. Neste caso, pode-se somente aproximar o tempo de espera usando-se a fórmula que refere a do sistema M/M/m, ou então, as fórmulas existentes para os limites superior e inferior do tempo de espera do sistema G/G/m.

No primeiro trecho do "paper", fórmulas para o limite superior e inferior são desenvolvidas e delas uma fórmula aproximada para o tempo de espera no sistema M/M/m tem origem. Estas fórmulas nao sao mais complicadas do que aquela do tempo de espera para o sistema M/M/1.

Estas resultados são extensivos para os sistemas M/G/m no segundo trecho do"paper". Mostra-se que estes resultados são muito melhores do que os obtidos da aproximaçao no sistema M/M/m e que os novos limites pertencem aos conhecidos no sistema G/G/m.

Palavras Chave

Sistema de fila com vários sevidores, tempo de espera, comprimento da fila, limites superior e inferior, aproximaçao.

Contents

1 - The Waiting Time and the Queue Length in M/G/m-Queueing Systems

In this chapter it is shown that we can obtain the waiting time in M/G/m-queueing systems easily when the probability

$P_{\geq m} = P[k \geq m] = P[\text{all servers are active}]$

k: number of customers in the system

m: number of servers

is known.

In general the mean waiting time in a M/G/m-queueing system is the sum of the service times of the waiting customers divided by m (because m customers can be served at one time) plus the average remaining service time:

$$W = W_o + \frac{N_q}{m} \bar{x}$$

with

W: mean waiting time
W_o: mean remaining service time
N_q: mean queue length
$\bar{x}$: mean service time in a server

Using Little's result

$$N_q = \lambda \cdot W$$

λ: arrival rate of the arriving customers

and introducing the utilization factor

$$\rho = \frac{\lambda \cdot \bar{x}}{m}$$

we get immediately from this equation

$$W = \frac{W_o}{1 - \rho} \qquad (1)$$

The remaining service time W_0 in an active single server system is $\overline{x^2}/(2\overline{x})$ [1] and therefore in a multiserver system $\overline{x^2}/(2m\overline{x})$ if all servers are active (probability $P_{\geq m}$). If at least one server is idle then $W_0 = 0$ and we get

$$W_0 = \frac{\overline{x^2}}{2 \cdot m \cdot \overline{x}} \cdot P_{\geq m}$$

and finally the waiting time as a function of $P_{\geq m}$

$$W = \frac{\overline{x^2}}{2 \cdot m \cdot \overline{x} \cdot (1 - \rho)} \cdot P_{\geq m}$$

or the queue length

$$N_q = \frac{\rho}{2 \cdot (1 - \rho)} \cdot \frac{\overline{x^2}}{\overline{x}^2} \cdot P_{\geq m} \qquad (2)$$

or with the variance $\sigma_x^2 = \overline{x^2} - \overline{x}^2$

$$N_q = \frac{\rho}{2 \cdot (1 - \rho)} \cdot \left(\frac{\sigma_x^2}{\overline{x}^2} + 1\right) \cdot P_{\geq m}$$

and the coefficient of variation $C_b = \sigma_x/\overline{x}$

$$N_q = \frac{\rho}{2\,(1 - \rho)} \cdot (C_b^2 + 1) \cdot P_{\geq m}$$

In the following chapters we only consider the queue length N_q; from this we can easily obtain both the waiting time W and the response time T using the wellknown formulas [1]

$$W = \frac{1}{\lambda} \cdot N_q \qquad \text{and} \qquad T = W + \overline{x}$$

The mean number in the system is also immediately available:

$$N = \lambda \cdot T$$

From this formula (2) we will proceed develop exact and approximate solutions and upper and lower bounds by considering the probability $P_{\geq m}$.

2 - The Queue Length in M/M/m-Queueing Systems

In the special case of a M/M/m-queueing system we can now obtain an exact result for the queue length. For exponential distribution we have $c_b = 1$ and therefore

$$N_q = \frac{\rho}{1 - \rho} \cdot P_{\geq m} \qquad (3)$$

With the known probability [1]

$$p_k = P[k \text{ customers in the system}] = \begin{cases} p_o \cdot \dfrac{(m\rho)^k}{k!} & k \leq m \\ p_o \cdot \dfrac{\rho^k \cdot m^m}{m!} & k \geq m \end{cases}$$

and

$$p_o = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1 - \rho)} \right]^{-1} \qquad (4)$$

we get

$$P_{\geq m} = \sum_{k=m}^{\infty} p_k = p_o \cdot \frac{(m\rho)^m}{m!(1 - \rho)} \qquad (5)$$

and finally the wellknown formula [2]:

$$N_q = \frac{m^m \cdot \rho^{m+1}}{m!(1 - \rho)^2} \cdot \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1 - \rho)} \right]^{-1} \qquad (6)$$

3 - Upper and Lower Bounds for the Queue Length in M/M/m-Queueing Systems

In many cases we only want to get an approximate value of the queue length N_q and then equation (5) is too expensive. Then it is advantageous to approximate $P_{\geq m}$ by an expression of the form ρ^n to find upper and lower bounds for N_q.

An upper bound for N_q we get with the definition of ρ:

ρ = E [fraction of active servers]

ρ = P [an arbitrary server is active]

and from this follows immediately:

$\rho \geq$ P[m servers are active] $= P_{\geq m}$

We can also show this by using the formula for $P_{\geq m}$ equation (5)

$$\rho \overset{?}{\geq} P_{\geq m} = p_0 \cdot \frac{(m\rho)^m}{m!(1-\rho)}$$

with equation (4)

$$\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{m^m \rho^m}{m!(1-\rho)} \overset{?}{\geq} \frac{m^m \rho^{m-1}}{m!(1-\rho)}$$

and two simplification steps:

$$\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} \overset{?}{\geq} \frac{m^m}{m!} \rho^{m-1} = \frac{m^{m-1}}{(m-1)!} \rho^{m-1}$$

$$\sum_{k=0}^{m-2} \frac{(m\rho)^k}{k!} \geq 0 \qquad \text{for} \quad \rho \geq 0 \quad \text{and} \quad m \geq 1$$

Similarly can be shown that

$$\rho^2 \geq P_{\geq m} \quad \text{for} \quad m \geq 4 \quad \text{and} \quad \rho^3 \geq P_{\geq m} \quad \text{for} \quad m \geq 7$$

but there is no simple general relation between m and the exponent n of the upper bound ρ^n.

We get a general lower bound if we consider m independent M/M/m-queueing systems which all have the same utilization factor ρ. Then we have

$$P[\text{an arbitrary server is active}] = \rho$$

$$P[\text{all servers are active}] = \rho^m$$

because of the independency of the systems and therefore for a single M/M/m-system

$$P_{\geq m} = P[\text{all servers are active}] \geq \rho^m$$

because in this case one or more servers can be idle only when no job is waiting. In the first case it is possible that a job is waiting and one or more servers are idle at the same time because each server has its own queue.

We can also verify this by using the formula for $P_{\geq m}$ equation (5)

$$\rho^m \overset{?}{\leq} p_0 \cdot \frac{(m\rho)^m}{m!(1-\rho)}$$

with equation (4)

$$\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \frac{1}{(1-\rho)} \overset{?}{\leq} \frac{m^m}{m!} \cdot \frac{1}{1-\rho}$$

$$\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} \overset{?}{\leq} \frac{m^m}{m!} \cdot \frac{1-\rho^m}{1-\rho} = \frac{m^m}{m!} \cdot \sum_{k=0}^{m-1} \rho^k$$

$$\frac{m^k}{k!} \cdot \rho^k \overset{?}{\leq} \frac{m^m}{m!} \rho^k$$

$$\frac{m^k}{k!} \leq \frac{m^m}{m!} \quad \text{for } m \geq k \text{ and } \rho \geq 0$$

ρ^m is in fact the best lower bound if we use only expressions of the form ρ^n with $n \in N$. This is true, because we can show that ρ^{m-1} is no general lower bound.

Proof:

$$\rho^{m-1} \gtrless P_{\geq m}$$

with equation (4) and (5)

$$\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{m^m \rho^m}{m!(1-\rho)} \overset{?}{\gtrless} \frac{m^m \rho}{m!(1-\rho)}$$

for $\rho \ll 1$ $\qquad 1 \geq 0$

We can get a better lower bound if we apply some restrictions to ρ (for example $\rho \geq 0,1$). But as in the case of the upper bound there is no simple relation which may be used to find the value of the exponent n of the lower bound ρ^n for a given restiction to ρ.

Our derived upper bound ρ and lower bound ρ^n have the advantage that they are general for all values $1 \geq \rho \geq 0$ and $m \geq 1$ without any restriction and that we get very simple formulas for the upper bound N_{qu} and lower bound N_{ql} of the queue length from equation (3)

$$N_{ql} = \frac{\rho^{m-1}}{1-\rho} \leq N_q \leq \frac{\rho^2}{1-\rho} = N_{qu} \qquad (7)$$

for $0 \leq \rho \leq 1$ and $1 \leq m \leq \infty$

Table 1 and figure 1,2 and 3 give the values of the bounds and the exact value of N_q as a function of ρ for m = 3; 5 and 10. We see from the table and the figures and also immediately from equation (7) that the relative deviation of the bounds from the exact value increases when m increases and ρ decreases.

For small values of ρ the relative deviations of the bounds are considerable but the absolut values of the queue length are very small in this region ($N_q \leq 0,5$ if $\rho \leq 0,5$ for all m). For relevant values of N_q (>1) our derived upper and lower bounds are good approximations if m is not too great.

4 - An Approximation for the Queue Length in M/M/m-Queueing Systems

We see from table 1 and especially from figure 1, 2 and that the exact value of N_q usually lies near the mean value of the upper and lower bound and it is only for small values that it approaches closer to the lower bound. Therefore we expect that the mean value of the upper and the lower bound is a good approximation formula for the queue length. That this is really the case we see from table 2 and the figures 1, 2 and 3 which show the values of the arithmetic and geometric means of the bounds. Both mean values are suprisingly good approximations for the exact value. The geometric mean value is closer to the lower bound and therefore better for small values of ρ, the arithmetic mean value is better for large values of ρ.

For the arithmetic mean value we get:

$$P_{\geq ma} = \frac{\rho + \rho^m}{2} \qquad \text{or} \qquad N_{qa} = \frac{\rho^2}{1-\rho} \cdot \frac{1+\rho^{m-1}}{2} \tag{8}$$

and for the geometric mean value:

$$P_{\geq mg} = \rho^{\frac{m+1}{2}} \qquad \text{or} \qquad N_{qg} = \frac{\rho^{\frac{m+3}{2}}}{1-\rho} = \frac{\rho^2}{1-\rho}\,\rho^{\frac{m-1}{2}} \tag{9}$$

These approximation formulas are much simpler than the exact formula equation (6) and yield surprisingly good results.

5 - Upper and Lower Bounds for the Queue Length in M/G/m-Queueing Systems

To find an upper and a lower bound for the queue length in M/G/m-queueing systems we can use equation (2) and substitute $P_{\geq m}$ in this formula by the upper and lower bound for this probability from chapter 3

$$\rho^m \leq P_{\geq m} \leq \rho$$

since the derivation for these bounds is also correct for M/G/m-systems. We obtain with equation (2)

$$N_{ql} = \frac{\rho^{m+1}}{2\cdot(1-\rho)} \cdot \frac{\overline{x^2}}{\bar{x}^2} \leq N_q \leq \frac{\rho^2}{2\,(1-\rho)} \cdot \frac{\overline{x^2}}{\bar{x}^2} = N_{qu} \qquad (10)$$

By comparing these bounds with the bounds for M/M/m-systems we find that we get the same expressions if we normalize the values for M/G/m-systems with $\overline{x^2}/(2\cdot\bar{x}^2)$. Therefore we can use table 1 and the bounds in figure 1, 2 and 3 as well if we multiply those values with $\overline{x^2}/(2\cdot\bar{x}^2)$. Table 3 shows the values of this factor for some wellknown distributions:

Distribution	$\overline{x^2}/(2\cdot\bar{x}^2)$
D	1/2
E_r	(r+1)/(2r)
M	1
H_R	>1

Table 3

We get another lower bound if we use the wellknown values of the state probability p_k for M/G/∞-systems.
We have

$$P_{\geq m\; M/G/m} \geq P_{\geq m\; M/G/\infty} \qquad (11)$$

because there are always more jobs in a M/G/m-system than in a M/G/∞-system if the mean value of the service time of a server and the arrivalrate are the same because of the finite number of servers in a M/G/m-system.
With [1]

$$p_{k\ M/G/\infty} = \frac{\lambda\bar{x}}{k!}\, e^{-\lambda\bar{x}}$$

we get

$$P_{\geq m\ M/G/\infty} = e^{-\lambda\bar{x}} \cdot \sum_{k=m}^{\infty} \frac{(\lambda\bar{x})^k}{k!}$$

$$= 1 - e^{-\lambda\bar{x}} \cdot \sum_{k=0}^{m-1} \frac{(\lambda\bar{x})^k}{k!}$$

or with $\rho = (\lambda\bar{x})/m$

$$P_{\geq m\ M/G/\infty} = 1 - e^{-m\rho} \cdot \sum_{k=0}^{m-1} \frac{(m\,\rho)^k}{k!}$$

and we obtain with this formula and equation (2) and (11)

$$N_{ql_2} = \frac{\rho}{2\cdot(1-\rho)} \cdot \frac{\overline{x^2}}{\bar{x}^2} \cdot \left(1 - e^{-m\rho} \cdot \sum_{k=0}^{m-1} \frac{(m\,\rho)^k}{k!}\right) \qquad (12)$$

Now it is advantageous to use as final lower bound the maximum of both the derived lower bounds for the instanteneous utilization ρ. $N_{ql,1}$ is greater for large values of ρ and $N_{ql,2}$ is greater for small values of ρ.

Therefore we get as final lower bound for the queue length in M/G/m-systems

$$N_{q_1} = \frac{\rho}{2\cdot(1-\rho)} \cdot \frac{\overline{x^2}}{\overline{x}^2} \cdot \max(\rho^m,\ 1 - e^{-m\rho} \sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!}) \qquad (13)$$

Figure 4 and 5 show both the lower bounds and the upper bound. We see from these figures that as expected the difference between the upper and the lower bound is greater for large m. But this is partially compensated for by the fact that for greater m the new lower bound is a greater improvement (for m=5 in the region $0 \le \rho \le 0,8$ and for m=10 in the region $0 \le \rho \le 0,9$).

6 - Comparison with the G/G/m Bounds

In this chapter we want to show that the new derived bounds for M/G/m-systems are better than the wellknown bounds for G/G/m-systems applied to M/G/m-systems.

For G/G/m-systems there exist the following upper and lower bounds for the waiting time [3,4,5]:

$$W_1 = \frac{\rho^2 C_b - \rho\cdot(2-\rho)}{2\cdot\lambda\cdot(1-\rho)} - \frac{\overline{x^2}\cdot(m-1)/m^2}{2\cdot\overline{x}} \le W$$

and

$$W_u = \frac{\sigma_a^2 + \sigma_b^2/m + \overline{x}^2\cdot(m-1)/m^2}{2\cdot\overline{t}\cdot(1-\rho)} \ge W$$

From this easily we get bounds for the queue length in M/G/m-systems:

$$N_{q_1}^* = \frac{\rho}{2\cdot(1-\rho)} \cdot \left(\frac{\overline{x^2}}{\overline{x}^2}(m\rho + 1 - m) - 2\right) \le N_q$$

and

$$N_{qu}^{*} = \frac{1 + \rho^2 (\overline{x^2}/\bar{x}^2 - 1)}{2(1 - \rho)} \geq N_q$$

Now we want to prove that

$$N_{ql}^{*} \leq N_{ql}$$

Proof:

$$\frac{\rho}{2\cdot(1 - \rho)} \cdot ((m\rho - m + 1) \cdot \frac{\overline{x^2}}{\bar{x}^2} - 2) \overset{?}{\leq} \frac{\rho^{m+1}}{2(1 - \rho)} \cdot \frac{\overline{x^2}}{\bar{x}^2}$$

or

$$\frac{\overline{x^2}}{\bar{x}^2} \cdot (m\rho - m + 1 - \rho^m) \overset{?}{\leq} 2$$

this is true also if

$$m \cdot (\rho - 1) + 1 - \rho^m \overset{?}{\leq} 0$$

or

$$m \geq \frac{1 - \rho^m}{1 - \rho} = \sum_{i=0}^{m-1} \rho^i$$

and this is true for $\rho \leq 1$.

For the upper bound we want to prove

$$N_{qu}^{*} \geq N_{qu}$$

Proof:

$$\frac{1 + \rho^2 \cdot (\overline{x^2}/\bar{x}^2 - 1)}{2\cdot(1 - \rho)} \overset{?}{\geq} \frac{\rho^2}{2\cdot(1 - \rho)} \cdot \frac{\overline{x^2}}{\bar{x}^2}$$

or

$$1 - \rho^2 \overset{?}{\geq} 0$$

$$1 \geq \rho^2$$

this is true also for $\rho \leq 1$.

7 - An Approximation of the Queue Length in M/G/m Queueing Systems

To get an approximation formula for the queue length we assume that the value of the queue length is always an average value between the upper and the lower bound (compare the case of M/M/m systems). Then we get an approximation formula as in chapter 4 using the arithmetic and the geometric mean values and obtain with equation (2) and (8):

$$N_{qa} = \frac{\rho}{2(1-\rho)} \cdot \frac{\overline{x^2}}{\overline{x}^2} \cdot \frac{\rho + \rho^m}{2}$$

and

$$N_{qg} = \frac{\rho}{2(1-\rho)} \cdot \frac{\overline{x^2}}{\overline{x}^2} \cdot \rho^{\frac{m+1}{2}}$$

We get a better mean value between the upper and the lower bound if we use the exact value of the probability $P_{\geq m}$ of M/M/m systems (see figure 1, 2 and 3):

$$P_{\geq m\ M/M/m} = p_0 \cdot \frac{(m\rho)^m}{m! \cdot (1-\rho)}$$

(p_0 see equation (4))

and from this we get the approximation formula:

$$N_q \approx \frac{\rho}{2\,(1-\rho)} \cdot \frac{\overline{x^2}}{\overline{x}^2} \cdot P_{\geq m \; M/M/m} \tag{14}$$

Comparing equation (14) with the formula for the queue length in M/M/m systems (equation (3)) which is often used as an approximation formula for the queue length in M/G/m systems we see that equation (14) is an improvement with the factor $\overline{x^2}/(2\cdot\overline{x}^2)$.

References

[1] Kleinrock, L.: Queueing Sysems, Volume 1: Theory, John Wiley & Sons, New York 1975

[2] Cobham, A.: Priority assignment in waiting line problems, Operation Research 3, 1955

[3] Kingman, J. F. C.: Inequalities in the Theory of Queues, Journal of the Royal Statistical Society, Series B, 32, 1970

[4] Brumelle, S. L.: Some Inequalities for Parallel Server Queues, Operation Research, 19, 1971

[5] Kleinrock, L.: Queueing Systems, Volume 2: Computer Applications, John Wiley & Sons, New York, 1976

ρ	Nqu	m=3 Nqex	Nql	m=5 Nqex	Nql	m=10 Nqex	Nql
0,1	0,01	0,00	0,00	-	-	-	-
0,2	0,05	0,01	0,00	0,00	0,00	-	-
0,3	0,13	0,03	0,01	0,01	0,00	-	-
0,4	0,27	0,09	0,04	0,04	0,01	0,00	0,00
0,5	0,50	0,24	0,13	0,13	0,03	0,04	0,00
0,6	0,90	0,53	0,32	0,35	0,12	0,15	0,01
0,7	1,61	1,14	0,80	0,88	0,39	0,52	0,06
0,8	3,20	2,56	2,05	2,22	1,31	1,64	0,42
0,9	8,10	7,29	6,56	6,86	5,31	6,02	3,13
0,95	18,05	17,23	16,15	16,68	14,70	15,67	11,36
0,99	98,01	97,12	96,03	96,52	94,15	95,44	89,50

Table 1 Upper and lower bounds and the exact value of the queue length in M/M/m queueing systems

ρ	m = 3 Nqex	Nqa	Nqg	m = 5 Nqex	Nqa	Nqg	m = 10 Nqex	Nqa	Nqg
0,1	0,00	0,01	0,00	0,00	0,01	0,00	-	-	-
0,2	0,01	0,03	0,01	0,00	0,02	0,00	-	-	-
0,3	0,03	0,07	0,04	0,01	0,06	0,01	0,00	0,06	0,00
0,4	0,09	0,16	0,11	0,04	0,14	0,04	0,00	0,13	0,00
0,5	0,24	0,31	0,25	0,13	0,26	0,13	0,04	0,25	0,02
0,6	0,53	0,61	0,54	0,35	0,51	0,32	0,15	0,46	0,09
0,7	1,14	1,22	1,14	0,88	1,01	0,80	0,52	0,85	0,33
0,8	2,56	2,62	2,56	2,22	2,26	2,05	1,64	1,81	1,18
0,9	7,29	7,33	7,29	6,86	6,71	6,56	6,02	5,62	5,04
0,95	17,23	17,18	17,15	16,68	16,38	16,29	15,67	14,71	14,33
0,99	97,12	97,12	97,02	96,52	96,03	96,06	95,44	93,75	93,67

Table 2 Approximate and exact values of the queue length in M/M/m queueing systems

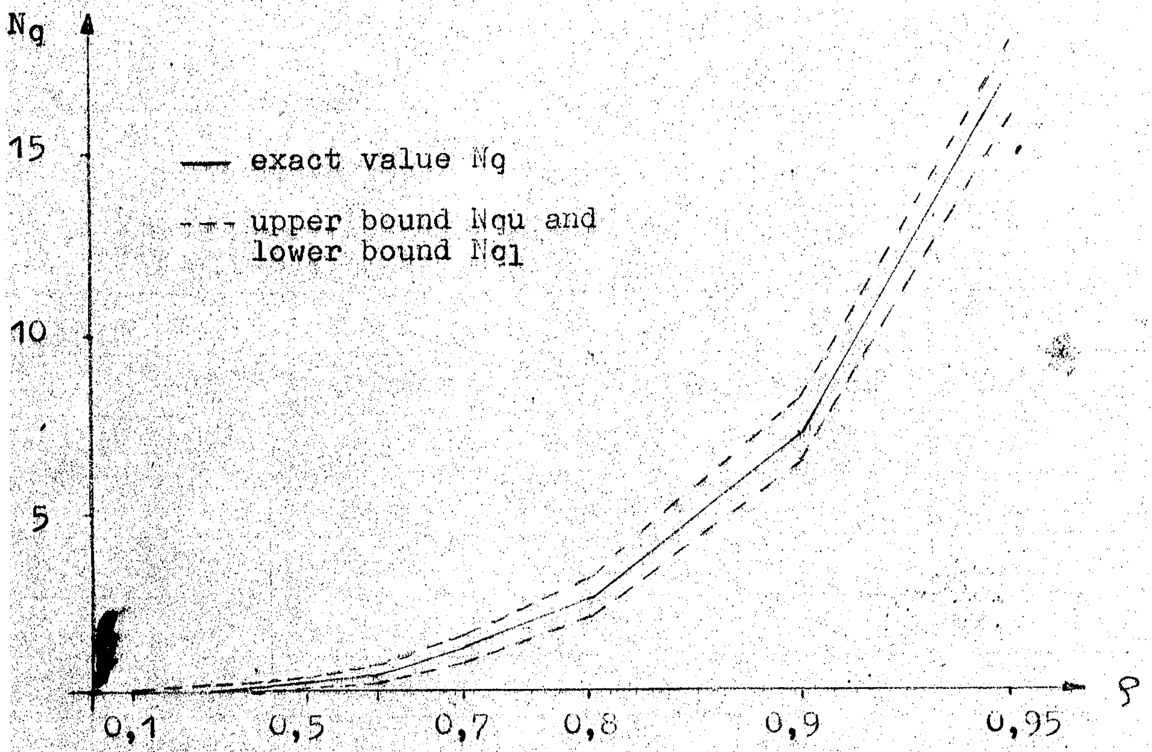

Figure 1 Queue length Nq in a M/M/3 queueing system

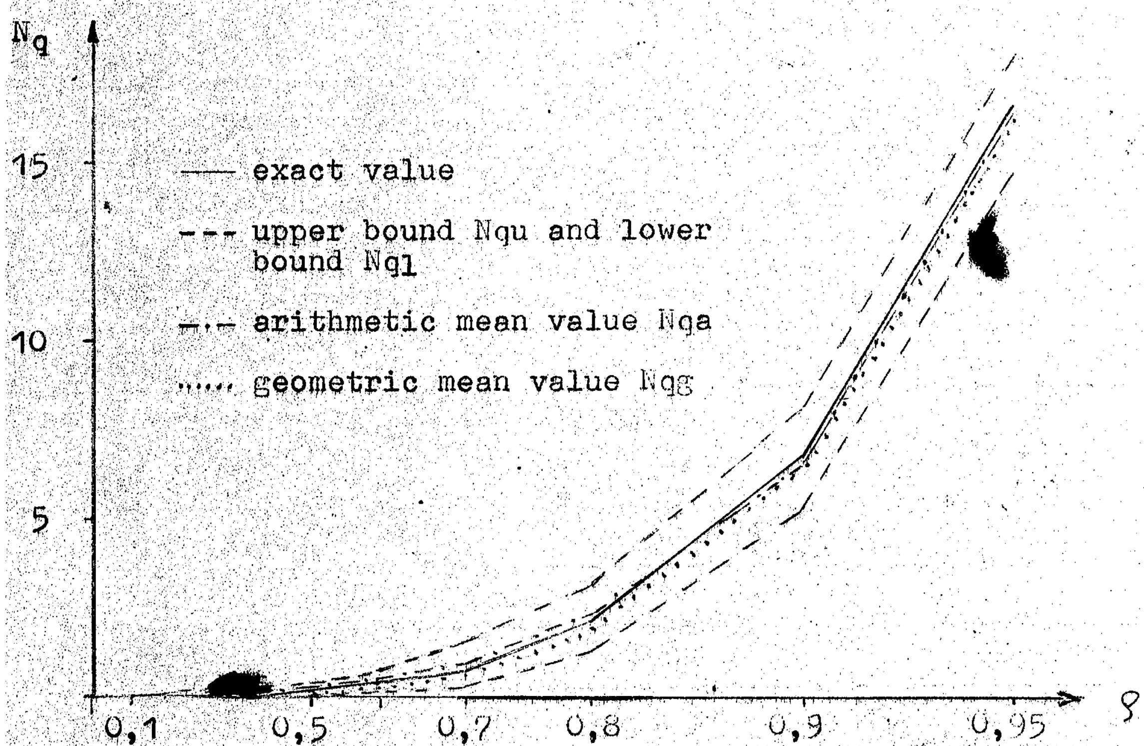

Figure 2 Queue length Nq in a M/M/5 queueing system

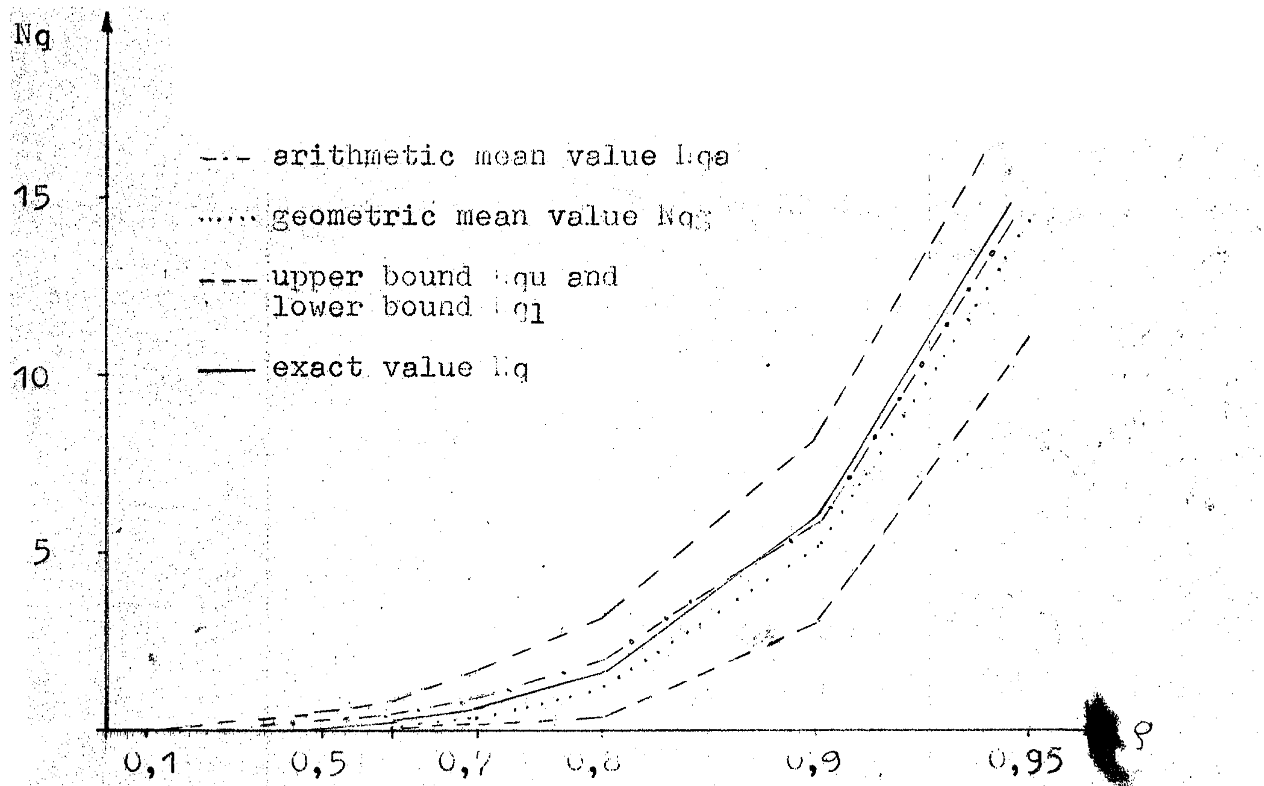

Figure 3 Queue length Nq in a M/M/10 queueing system

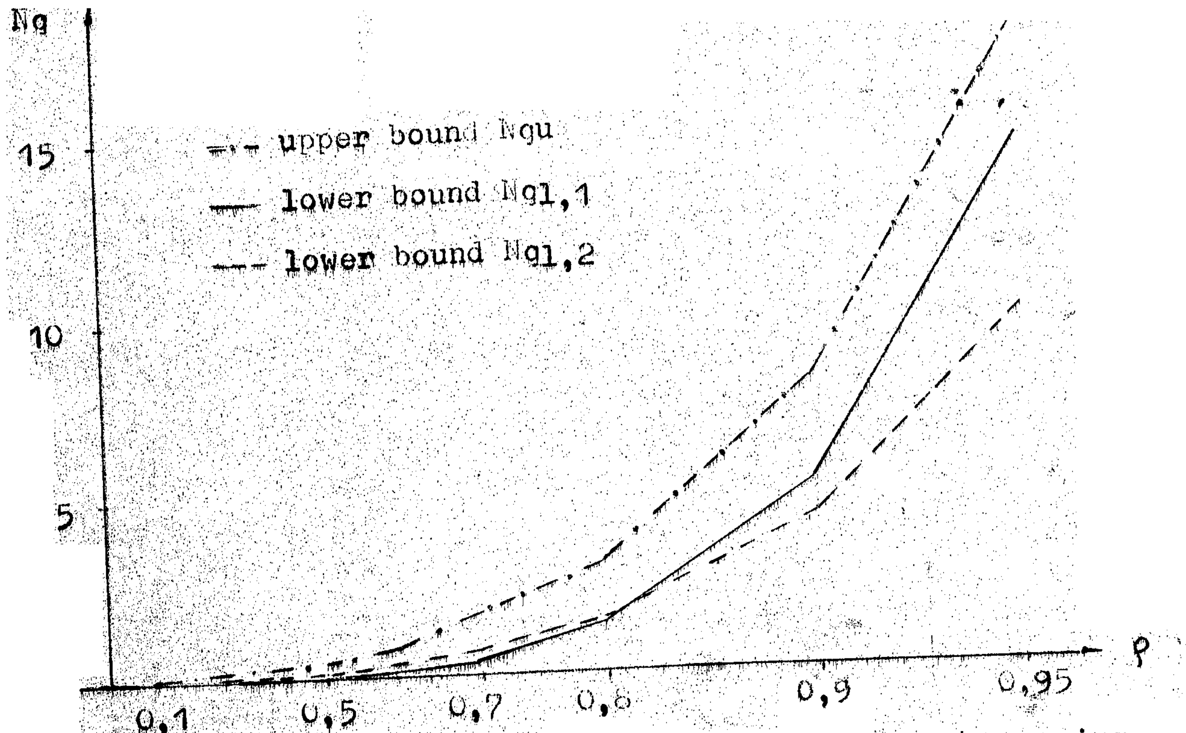

Figure 4 Bounds for the queue length in M/G/5 queueing systems

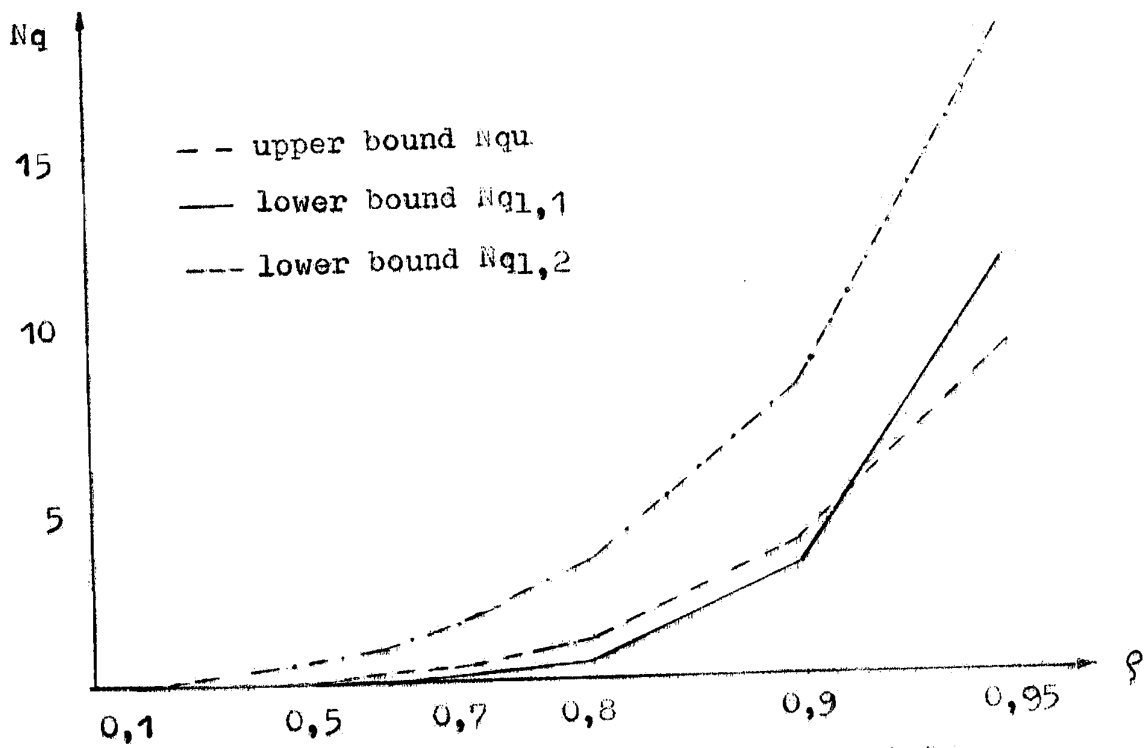

Figure 5 Bounds for the queue length in M/G/10 queueing systems