Printing Requisition / Graphic Services

569

Dept. No. 09854

Title or Description
Response Time Dist. of the M/M/m/N Queueing Model / J.W.Wong

Date 4/5/78

Date Required A.S.A.P. PLEASE & THANK YOU

Account 126-6356-41

Signature

Signing Authority Johnny W. Wong

Department Computer Science

Room 5180

Phone 3143

Delivery
☐ Mail ☐ Via Stores
☐ Pick-up ☐ Other

1. Please complete unshaded areas on form as applicable. (4 part no carbon required).
2. Distribute copies as follows: White, Canary and Pink—Printing, Arts Library or applicable Copy Centre Goldenrod—**Retain.**
3. On completion of order, pink copy will be returned with printed material. Canary copy will be costed and returned to requisitioner, **Retain as a record of your charges.**
4. Please direct enquiries, quoting requisition number, to Printing/Graphic Services, Extension 3451.

Reproduction Requirements
☐ Offset ☐ Signs/Repro's ☐ Xerox

Number of Pages 16

Number of Copies 50

Type of Paper Stock
☑ Bond ☐ Book ☐ Cover ☐ Bristol ☐ Supplied

Paper Size
☐ 8½ x 11 ☐ 8½ x 14 ☐ 11 x 17

Paper Colour
☐ White ☐ Other

Ink
☐ Black

Printing
☐ 1 Side ☐ 2 Sides

Numbering to

Binding/Finishing Operations
☐ Collating ☑ Corner Stitching ☐ 3 Ring ☐ Tape ☐ Plastic Ring ☐ Perforating

Folding
Finished Size 8½ x 11

Cutting
Finished Size

Special Instructions
50 BACKS
50 COVERS
ENCLOSED.

Cost: Time/Materials		Fun.	Prod.Un.	Prod. Cl	Opr. No.	Mins.	Total	
Signs/Repro's	1							
Camera	2							
Correcting & Masking Negatives	3							
Platemaking	4			2	B			
Printing	5							
Bindery	6			2	B	3	25	
2 B.								
Sub. Total Time								

Film Qty	Size	**Plates** Qty	Size & Type	**Sub. Total Materials**	
Paper Qty	Size	**Plastic Rings** Qty	Size	**Prov. Tax**	
Outside Services				**Total**	

Graphic Services
May 68 482-2

Printing Requisition/Graphic Services

Doc. No. 37783

Title or Description CS-78-25

Date [illegible] **Date Required** [illegible] **Account** 126-6356-41

Signature E. Huang **Signing Authority** Johnny W. Huang

Department **Room** 175 **Phone** 111 **Delivery** ☑ Mail ☑ Pick-up ☐ Via Stores ☐ Other

1. Please complete unshaded areas on form as applicable. (4 part no carbon required).
2. Distribute copies as follows: White, Canary and Pink—Printing, Arts Library or applicable Copy Centre Goldenrod—**Retain.**
3. On completion of order, pink copy will be returned with printed material. Canary copy will be costed and returned to requisitioner, **Retain as a record of your charges.**
4. Please direct enquiries, quoting requisition number, to Printing/Graphic Services, Extension 3451.

Reproduction Requirements ☐ Offset ☐ Signs/Repro's ☐ Xerox **Number of Pages** 16 **Number of Copies** 50

Type of Paper Stock ☑ Bond ☐ Book ☐ Cover ☐ Bristol ☐ Supplied

Paper Size ☑ 8½ x 11 ☐ 8½ x 14 ☐ 11 x 17

Paper Colour ☐ White ☐ Other **Ink** ☑ Black

Printing ☐ 1 Side ☑ 2 Sides **Numbering** to

Binding/Finishing Operations ☑ Collating ☐ Corner Stitching ☐ 3 Ring ☐ Tape ☐ Plastic Ring ☐ Perforating

Folding Finished Size [illegible] **Cutting** Finished Size

Special Instructions

2 [illegible]

Cost: Time/Materials	Fun.	Prod.Un. Cl.	Prod.Opr. No.	Mins.	Total
Signs/Repro's	1				
Camera	2			1	
Correcting & Masking Negatives	3				
Platemaking	4				
Printing	5				
Bindery	6		22-20		
Sub. Total Time					
Sub. Total Materials					
Prov. Tax					
Total					

Film Qty	Size	**Plates** Qty	Size & Type
Paper Qty	Size	**Plastic Rings** Qty	Size
Outside Services			

Graphic Services
Mav 68 482-2

RESPONSE TIME DISTRIBUTION OF THE* M/M/m/N QUEUEING MODEL

by

J. W. Wong

Research Report CS-78-25

Department of Computer Science and
Computer Communications Networks Group

University of Waterloo
Waterloo, Ontario, Canada

---

* To appear in Operations Research.

A B S T R A C T

The M/M/m/N queueing model with a first-come, first-served discipline is analysed. It is shown that the distribution of response time (i.e., queueing time + service time) is asymptotically normal as N increases. The mean and variance of this normal distribution are both simple functions of the model parameters.

The model under consideration is shown in Figure 1. It represents N customers requesting service from a system of m parallel servers. Each customer is going through repeated periods of independent activity, waiting in queue, and receiving service from one of the m servers. The duration of independent activity and the service time are both exponential, with means $1/\lambda$ and $1/\mu$ respectively.

Sekino [4,5] has derived the distribution of response time (i.e., queueing time + service time) for the case of a first-come, first-served (FCFS) discipline. His result is rather complicated and the general shape of the distribution is not clearly shown. In this note, we show that the response time under FCFS is asymptotically normal as N increases. The mean and variance of this normal distribution are both simple functions of the model parameters. These results allow us to trivially characterize the response time at large N, and to efficiently evaluate the sensitivity of its mean and variance to changes in model parameters.

The M/M/m/N model finds applications in machine-repairman problems where a set of N machines is maintained by a pool of m repairmen. The response time corresponds to the elapsed time from the failure of a machine to the completion of repair of this machine. The M/M/m/N model has also been shown to give accurate predictions to the response time of existing time-sharing computer systems. Typical examples are Sekino's work on Multics [4,5], Scherr's work on the Compatible Time-Sharing System [3],

and Lassettre and Scherr's work on the IBM 360 Time-Sharing Option [6].

In a related work, Iglehart [7] has considered the M/M/m/N model with m = cN where c is a constant. He showed that when N is large, the number of customers in queue or in service is approximately normal.

2. Sekino's Result [4,5]

Let P(n), n = 0,1,...,N, be the equilibrium probability that there are n customers in queue or in service. For the M/M/m/N model, the solution to P(n) can be found in Saaty [2]:

$$P(n) = \begin{cases} \{N!/[(N-n)!n!]\}\ \rho^n\ P(0) & 0 \le n \le m \\ \{N!/[(N-n)!m!m^{n-m}]\}\ \rho^n\ P(0) & m \le n \le N \end{cases} \qquad (1)$$

where $\rho = \lambda/(m\mu)$ and

$$P(0) = \left[\sum_{n=0}^{m-1} \{N!/[(N-n)!n!]\}\ \rho^n + \sum_{n=m}^{N} \{N!/[(N-n)!m!m^{n-m}]\}\ \rho^n \right]^{-1}$$

Using P(n), Sekino [4,5] derived the following expression for A(n), the probability that an arriving customer (i.e., a customer entering the queue at the end of independent activity) finds n other customers in queue or in service:

$$A(n) = (N-n)\lambda P(n) \Big/ \sum_{n=0}^{N-1} (N-n)\lambda P(n) = (N-n)P(n)/\bar{q} \qquad (2)$$

where $\bar{q}$ is the mean number of customers in independent activity.

Let $T(x)$ be the response time distribution and $T^*(s)$ be the Laplace transform of its probability density function, i.e.,

$$T^*(s) = \int_0^\infty e^{-sx} dT(x)$$

$T^*(s)$ can be written as:

$$T^*(s) = \sum_{n=0}^{N-1} A(n)\, T^*(s|n) \tag{3}$$

where $T^*(s|n)$ is $T^*(s)$ conditioned on an arriving customer finding n other customers in queue or in service. Since the service time distribution is exponential and the queueing discipline is FCFS, we have:

$$T^*(s|n) = \begin{cases} \mu/(s+\mu) & 0 \le n < m \\ [m\mu/(s+m\mu)]^{n-m+1}[\mu/(s+\mu)] & m \le n < N \end{cases} \tag{4}$$

Substituting Eq.(4) into Eq.(3), we get:

$$T^*(s) = [\mu/(s+\mu)]\left[\sum_{n=0}^{m-1} A(n) + \sum_{n=m}^{N-1} A(n)[m\mu/(s+m\mu)]^{n-m+1}\right] \tag{5}$$

$T^*(s)$ can be inverted directly to give:

$$T(x) = \sum_{n=0}^{m-1} A(n)(1-e^{-\mu x}) + \sum_{n=m}^{N-1} A(n)\left\{[m/(m-1)]^{n-m+1}(1-e^{-\mu x})\right.$$

$$\left. - (1/m)\sum_{j=0}^{n-m}[m/(m-1)]^{n-m+1-j}[1 - e^{-m\mu x}\sum_{k=0}^{j}(m\mu x)^k/k!]\right\}$$

The mean $\bar{T}$ and variance σ_T^2 can also be obtained from $T^*(s)$ by differentiation. They are given by:

$$\bar{T} = 1/\mu + \sum_{n=m}^{N-1} A(n)(n-m+1)/(m\mu)$$

$$\sigma_T^2 = 2/\mu^2 + \sum_{n=m}^{N-1} A(n)(n-m+1)(n+m+2)/(m\mu)^2 - \bar{T}^2$$

3. Normal Distribution at Large N

We now consider the model at large N. It can be shown from Eq.(1) that as N increases, P(0), P(1), ..., P(m-2), and P(m-1) all approach zero. In other words, the m servers are almost 100% busy at all time. This implies that when N is large, the box labelled "independent activity" approaches an $M/M/\infty$ model [2] with mean interarrival time and mean service time equal to $1/m\mu$ and $1/\lambda$ respectively. Using the $M/M/\infty$ model as an approximation, we get the following expression for Q(n), the equilibrium probability of having n customers in independent activity [2]:

$$Q(n) = (m\mu/\lambda)^n e^{-m\mu/\lambda}/n!$$

The mean number of customers in independent activity is then given by:

$$\bar{q} = m\mu/\lambda \qquad (6)$$

Substituting Eq.(6) into Eq.(2), and recognizing that $P(n) = Q(N-n)$, we get:

$$A(n) = (m\mu/\lambda)^{N-n-1} e^{-m\mu/\lambda}/(N-n-1)! \qquad (7)$$

It is easy to see from Eq.(7) that when N is large, $A(0), A(1), \ldots, A(m-2)$, and $A(m-1)$ also approach zero. Assuming that these probabilities are zero, we get from Eq.(5) and (7) the following approximation for $T^*(s)$:

$$T^*(s) = e^{-m\mu/\lambda}[\mu/(s+\mu)][m\mu/(s+m\mu)]^{N-m} \sum_{n=m}^{N-1} [(s+m\mu)/\lambda]^{N-n-1}/(N-n-1)!$$

When N is large, the sum on the right hand side approaches $e^{(s+m\mu)/\lambda}$. This implies that:

$$T^*(s) \rightarrow e^{s/\lambda}[\mu/(s+\mu)][m\mu/(s+m\mu)]^{N-m} \qquad (8)$$

We observe from Eq.(8) that the response time at large N is approximately given by the sum of N-m+2 independent random variables: the first is deterministic with mean $-1/\lambda$, the next is exponential with mean $1/\mu$, and each of the remaining N-m is exponential with mean $1/m\mu$. By the central limit theorem [1], the response time distribution approaches a normal distribution when N is large. The mean and variance of this distribution are

given by:

$$\bar{T} = N/(m\mu) - 1/\lambda \tag{9}$$

$$\sigma_T^2 = 1/\mu^2 + (N-m)/(m\mu)^2 \tag{10}$$

T(x) can therefore be written as:

$$T(x) = \int_0^x (\sigma_T\sqrt{2\pi})^{-1}\exp[-(t-\bar{T})^2/(2\sigma_T^2)]\, dt$$

We note from Eq.(9) and (10) that $\bar{T}$ and σ_T^2 are both simple explicit functions of the model parameters.

4. Numerical Examples

In this section, numerical examples are presented to compare our normal approximation with the exact analysis of Sekino [4,5]. In our first example, the parameter values are:

$$m = 1, \quad \lambda = 0.1, \quad \mu = 1.0$$

The mean and variance of response time are plotted against N in Figure 2. We observe that there is practically no difference between the exact analysis and our approximation when N is large.

The response time distribution for N = 20, 30, 40, and 50 are plotted in Figure 3. We observe that our normal approximation is reasonably accurate, but not as good as those for $\bar{T}$ and σ_T^2. The reason is that the central limit theorem is not required to get the approximate expressions for $\bar{T}$ and σ_T^2.

In our second example, the parameter values are:

$m = 4, \quad \lambda = 0.1, \quad \mu = 0.25$

This is equivalent to replacing the single server in the first example by four servers, each with 1/4 the capacity. The results are plotted in Figures 4 and 5. We have essentially the same observations as those made in the first example.

5. Discussion of Results

Since the derivation of the approximate expressions for $\bar{T}$, σ_T^2, and $T(x)$ are based on the condition that N is large, the applicability of these expressions will depend on how large N is. We know from Eq.(8) that N must be larger than $m\mu/\lambda$ for our approximation to be meaningful. Using $m\mu/\lambda$ as a convenient reference point, numerical computations have indicated that the approximation is accurate when $N \geq 3*(m\mu/\lambda)+m$. This observation is supported by the numerical examples presented in the last section. $3*(m\mu/\lambda)+m$ can therefore be used as a conservative threshold value of N above which the approximation is accurate.

As mentioned earlier, the M/M/m/N model has been shown to give accurate predictions to the response time of existing time-sharing computer systems [3,4,5]. The case of large N corresponds to the situation of heavy load. Our results indicate that at heavy load, the response time distribution is approximately normal, and the mean and variance of response time are both linear functions of N.

References

1. Papoulis, A. "Probability, Random Variables, and Stochastic Processes," McGraw-Hill, New York, 1965.
2. Saaty, T. "Elements of Queueing Theory," McGraw-Hill, New York, 1961.
3. Scherr, A. "An Analysis of Time-Shared Computer Systems," Project MAC, Massachusetts Institute of Technology, Cambridge, Massachusetts, MAC-TR-18, June 1965.
4. Sekino, A. "Performance Evaluation of Multiprogrammed Time-Shared Computer System," Project MAC, Massachusetts Institute of Technology, Cambridge, Massachusetts, MAC-TR-103, September 1972.
5. Sekino, A. "Response Time Distribution of Multiprogrammed Time-Shared Computer Systems," Proc. 6th Annual Princeton Conference on Information Sciences and Systems, 1972, pp. 613-619.
6. Lassettre, E. and Scherr, A. "Modeling the Performance of the OS/360 Time-Sharing Option (TSO)," Statistical Computer Performance Evaluation, edited by W. Freiberger, Academic Press, New York, 1972, pp. 57-72.
7. Iglehart, D. "Limiting Diffusion Approximations for the Many Server Queue and the Repairman Problem," Journal of Applied Probability, Vol. 2, 1965, pp. 429-441.

Acknowledgement

This work was supported in part by grants from the National Research Council of Canada and the University of Waterloo.

Figure 1. The M/M/m/N Model

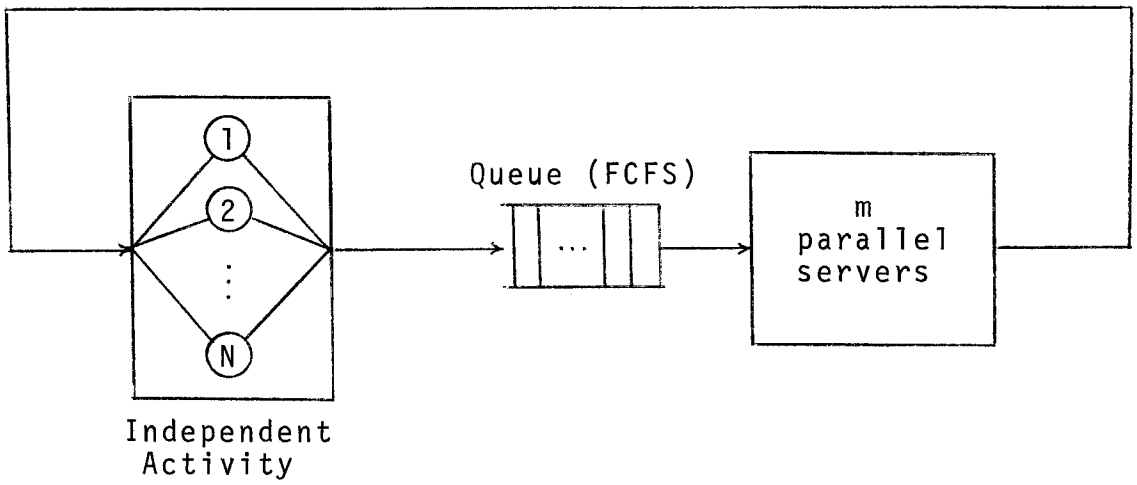
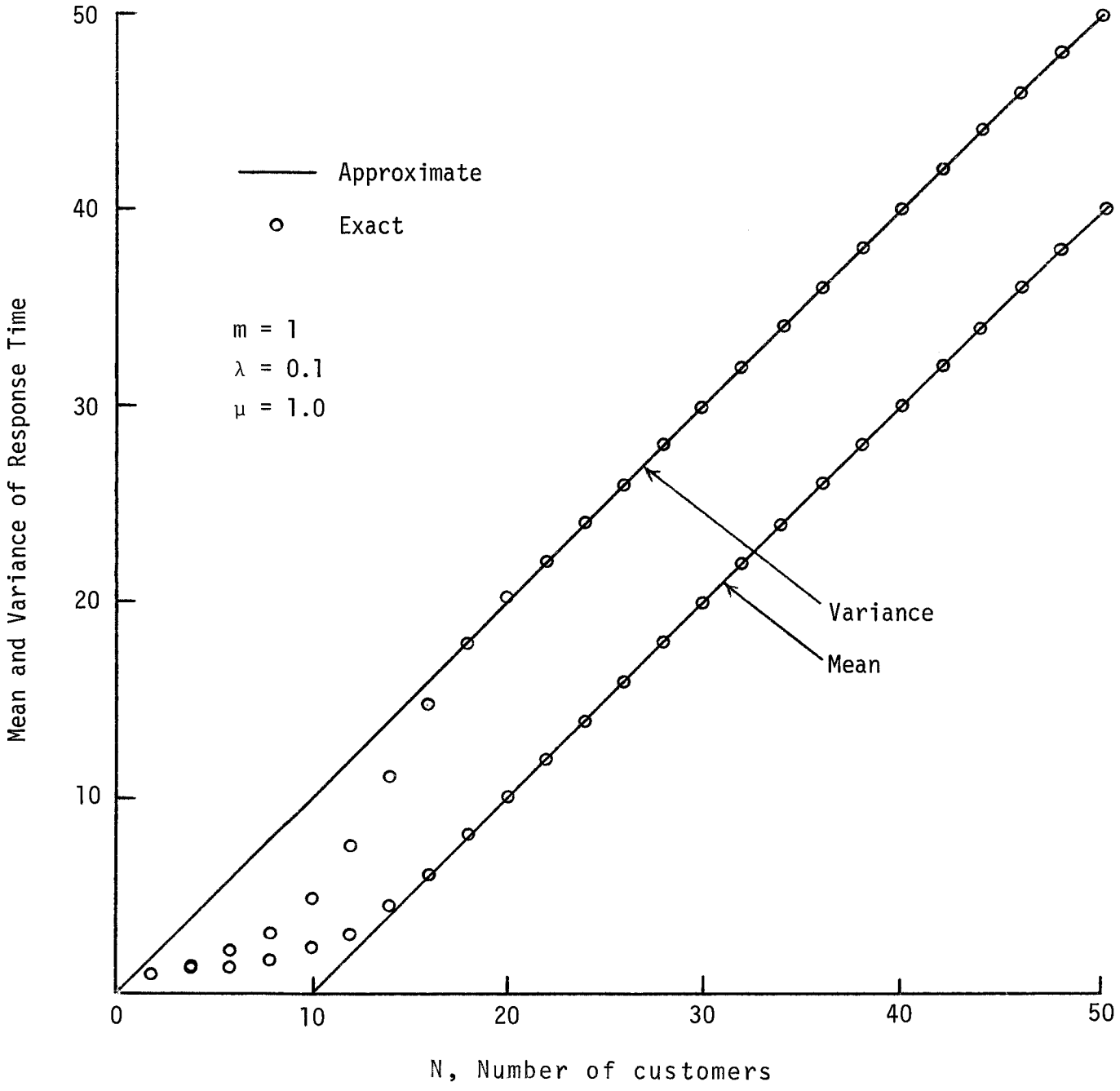


Figure 2. Mean and Variance of Response Time vs N



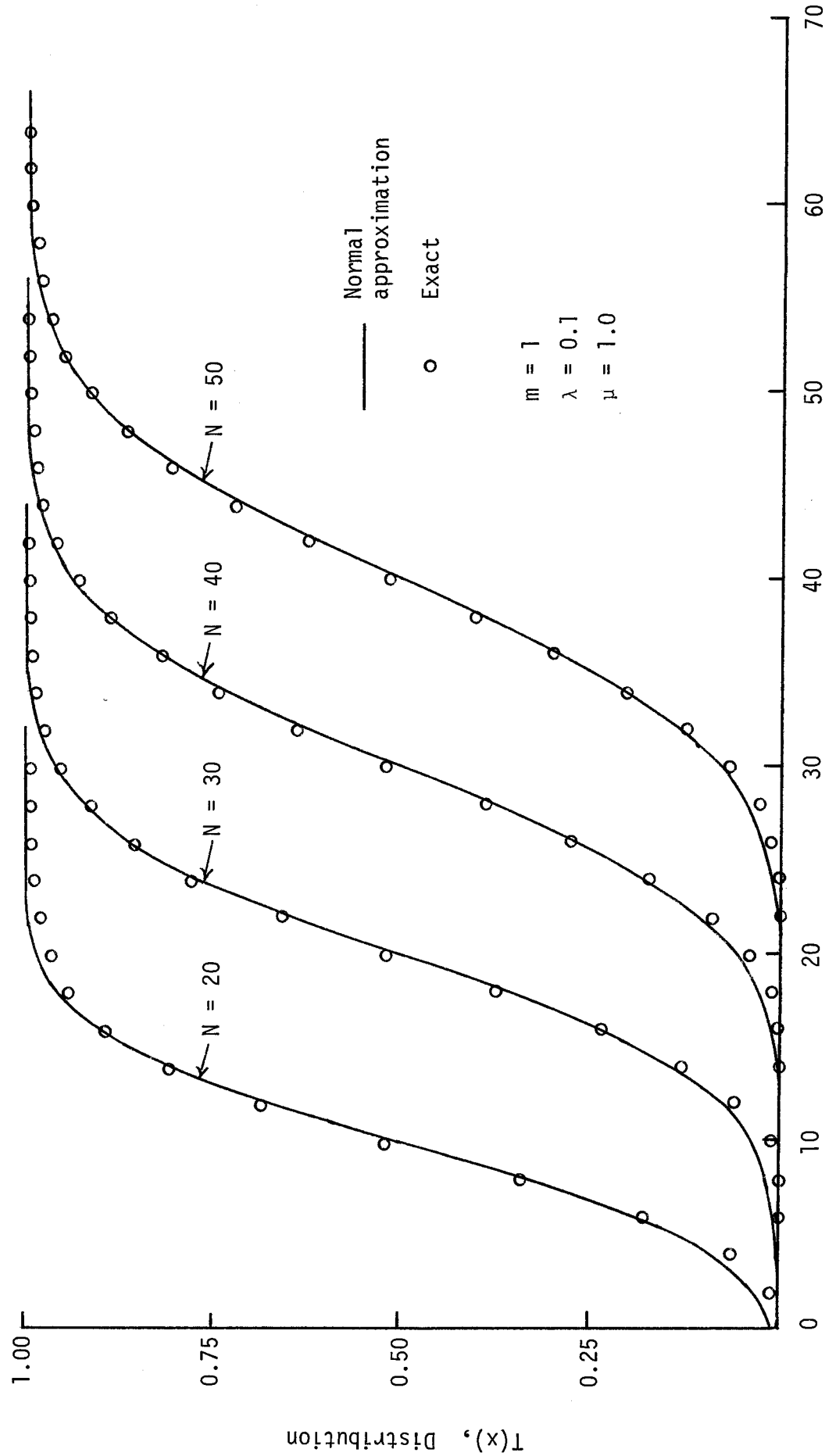
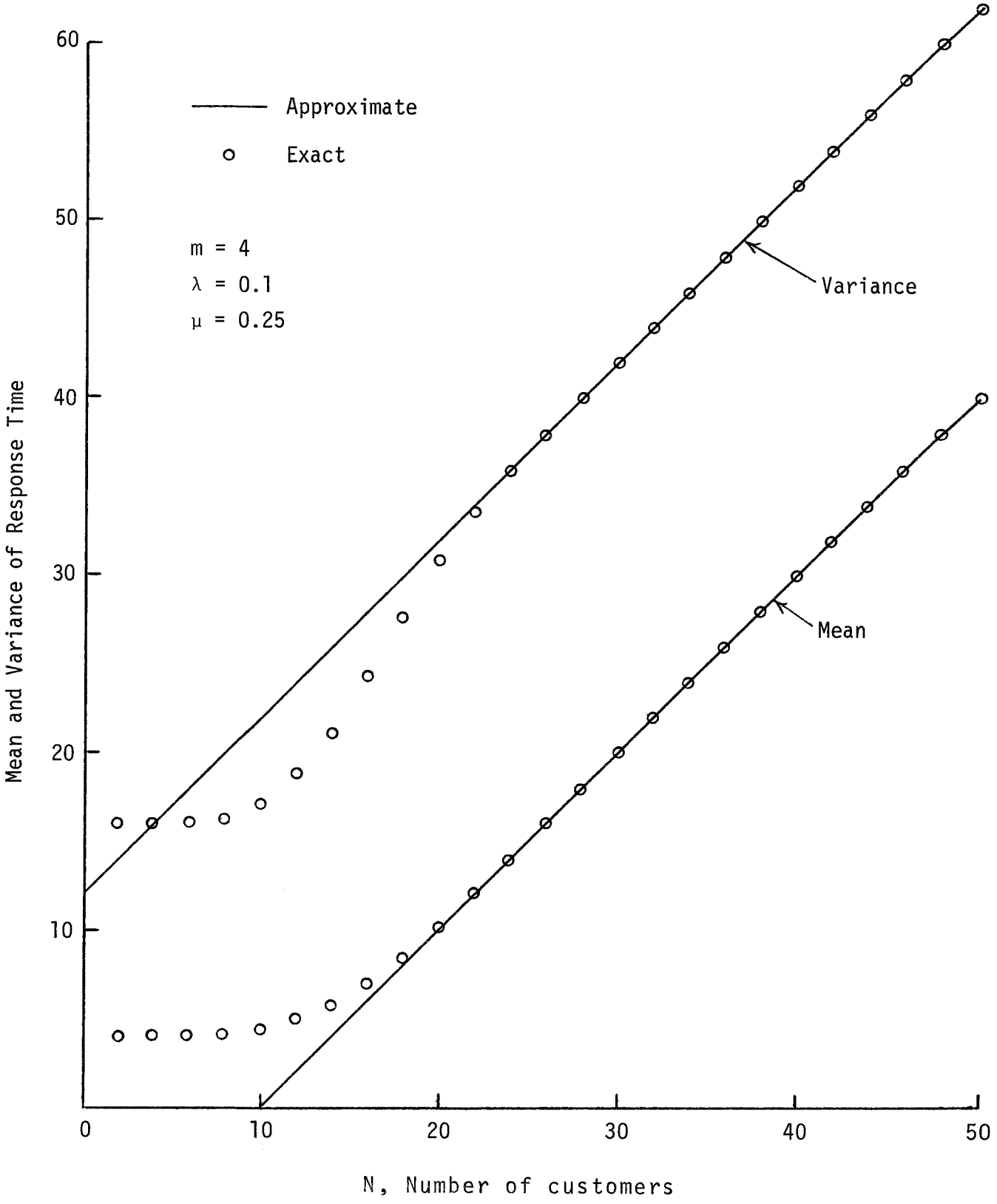


Figure 3. Response Time Distribution

Figure 4. Mean and Variance of Response Time vs N



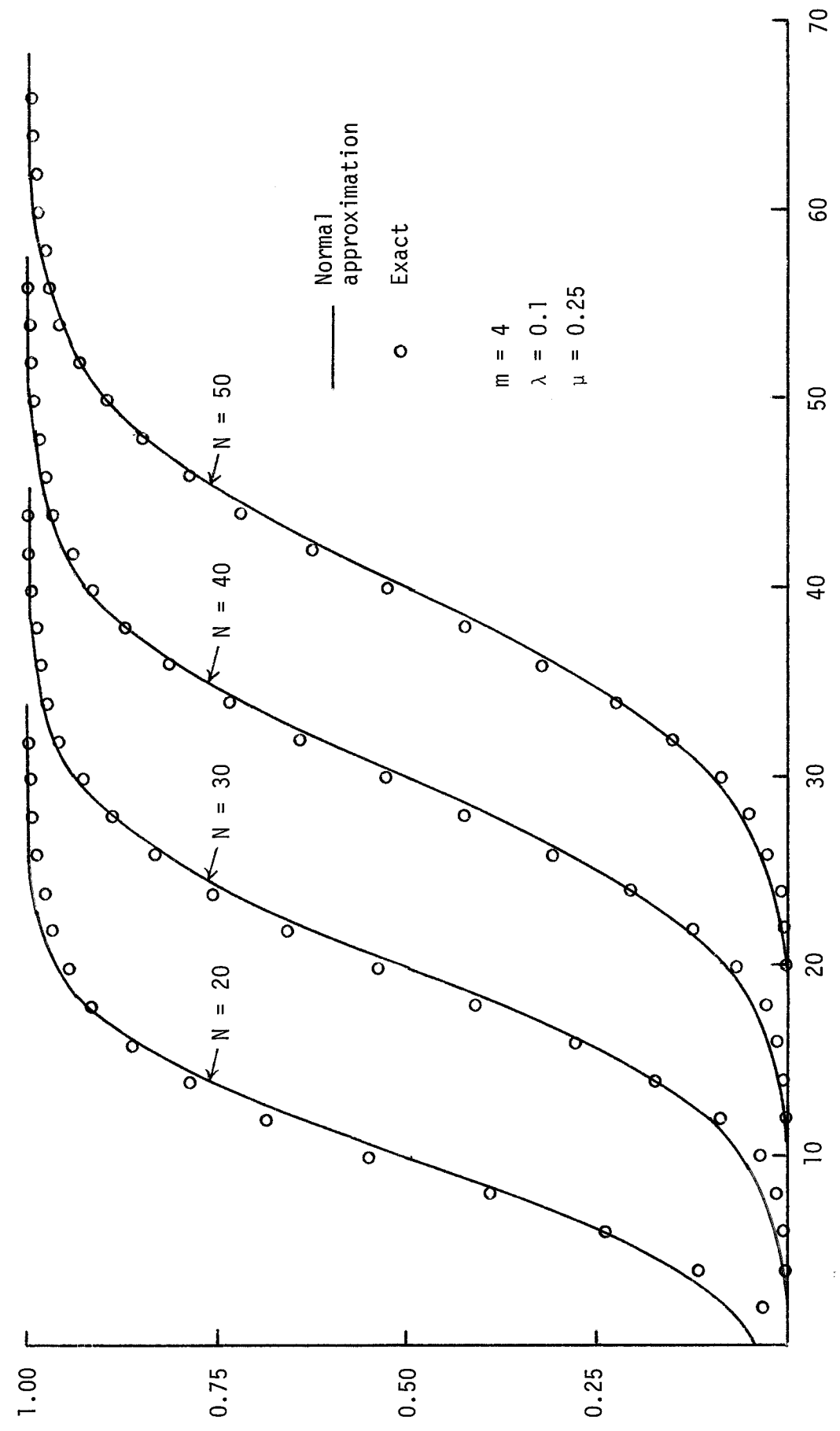


Figure 5. Response Time Distribution