



Operations Research

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Technical Note—Response Time Distribution of the M/M/m/N Queueing Model

J. W. Wong,

To cite this article:

J. W. Wong, (1979) Technical Note—Response Time Distribution of the M/M/m/N Queueing Model. Operations Research 27(6):1196-1202. <https://doi.org/10.1287/opre.27.6.1196>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

© 1979 INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes. For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

20. H. THEIL, "On the Use of Information Theory Concepts in the Analysis of Financial Statements," *Management Sci.* **15**, 459-480 (1969).
21. M. TRIBUS, *Rational Descriptions, Decisions, and Designs*, Pergamon, Elmsford, N.Y., 1969.
22. A. WIDAWSKY, "Development of Harbor Oil Spill Removal-Recovery Systems: Phase I," in *Proc. of Joint Conf. on Prev. and Contr. of Oil Spills*, pp. 379-386, American Petroleum Institute, Washington, D.C., 1975.
23. A. G. WILSON, "Further Developments of Entropy Maximizing Transport Models," *Transpn. Plan. Tech.* **1**, 183-193 (1973).
24. C. J. WILSON AND G. C. PHILIPPATOS, "Entropy, Market Risk, and the Selection of Efficient Portfolios," *Appl. Econ.* **4**, 209-220 (1972).
25. J. E. WILSON, P. F. NADEAU, AND J. S. PREMACK, "Navy Harbor Oil Spill Cleanup: A Progress Report," in *Proc. of Joint Conf. on Prev. and Contr. of Oil Spills*, pp. 373-378, American Petroleum Institute, Washington, D.C., 1975.

Response Time Distribution of the $M/M/m/N$ Queueing Model

J. W. WONG

University of Waterloo, Waterloo, Ontario

(Received April 1977; accepted April 1978)

We analyze the $M/M/m/N$ queueing model with a first-come, first-served discipline and show 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 [6, 7] 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 paper, 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 [6, 7], Scherr's work on the Compatible Time-Sharing System [5], and Lassetre and Scherr's work on the IBM 360 Time-Sharing Option [2].

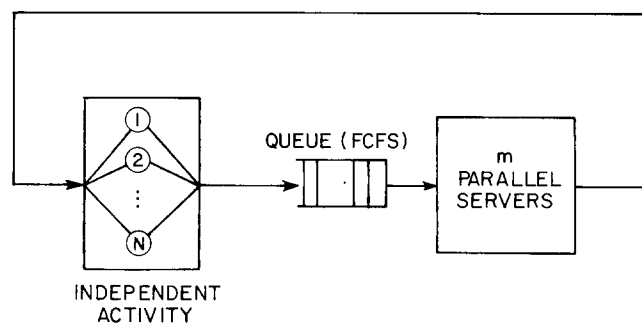


Figure 1. The $M/M/m/N$ model.

In a related work, Iglehart [1] 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.

1. SEKINO'S RESULT [6, 7]

Let $P(n)$, $n = 0, 1, \dots, N$, 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 [4]:

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

where $\rho = \lambda/(m\mu)$, and $P(0) = [\sum_{n=0}^{m-1} \{N!/[(N-n)!n!]\} \rho^n + \sum_{n=m}^N \{N!/[(N-n)!m!m^{n-m}]\} \rho^n]^{-1}$. Using $P(n)$, Sekino [6,7] 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:

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

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., $T^*(s) = \int_0^\infty \exp(-sx) dT(x)$. $T^*(s)$ can be written as:

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

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:

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

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

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

$T^*(s)$ can be inverted directly to give: $T(x) = \sum_{n=0}^{m-1} A(n)(1 - \exp(-\mu x)) + \sum_{n=m}^{N-1} A(n) \{ [m/(m-1)]^{n-m+1} (1 - \exp(-\mu x)) - (1/m) \sum_{j=0}^{n-m} [m/(m-1)]^{n-m+1-j} [1 - \exp(-m\mu x)] \sum_{k=0}^j (m\mu x)^k / k! \}$. The mean $\bar{T}$ and variance σ_T^2 can also be obtained from $T^*(s)$ by differentiation. They are given by: $\bar{T} = 1/\mu + \sum_{n=m}^{N-1} A(n)(n-m+1)/(m\mu)$, $\sigma_T^2 = 2/\mu^2 + \sum_{n=m}^{N-1} A(n)(n-m+1)(n-m+2)/(m\mu)^2 - \bar{T}^2$.

2. NORMAL DISTRIBUTION AT LARGE N

We now consider the model at large N . It can be shown from (1) that as N increases, $P(0)$, $P(1) \dots$, $P(m-2)$ and $P(m-1)$ all approach zero. In other words, the m servers are almost 100% busy at all times. This implies that when N is large, the part of the system labelled "independent activity" is well represented by an $M/M/\infty$ model [4] 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 [4]: $Q(n) = (m\mu/\lambda)^n \exp(-m\mu/\lambda)/n!$. The mean number of customers in independent activity is then given by:

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

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

$$A(n) = (m\mu/\lambda)^{N-n-1} \exp(-m\mu/\lambda)/(N-n-1)!. \quad (7)$$

It is easy to see from (7) that when N is large, $A(0)$, $A(1)$, $\dots$, $A(m-2)$, and $A(m-1)$ also approach zero. Assuming that these probabilities are zero, we get from (5) and (7) the following approximation for $T^*(s)$: $T^*(s) = \exp(-m\mu/\lambda) [\mu/(s + \mu)] [m\mu/(s + m\mu)]^{N-m} \sum_{n=m}^{N-1} [(s + m\mu)/\lambda]^{N-n-1}/(N-n-1)!$. When N is large, the sum on the right hand side approaches

$\exp((s + m\mu)/\lambda)$. This implies that:

$$T^*(s) \approx \exp(s/\lambda) [\mu/(s + \mu)] [m\mu/(s + m\mu)]^{N-m}. \quad (8)$$

We observe from (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 [3], the response time distribution approaches a normal distribution when N is large. The mean and variance

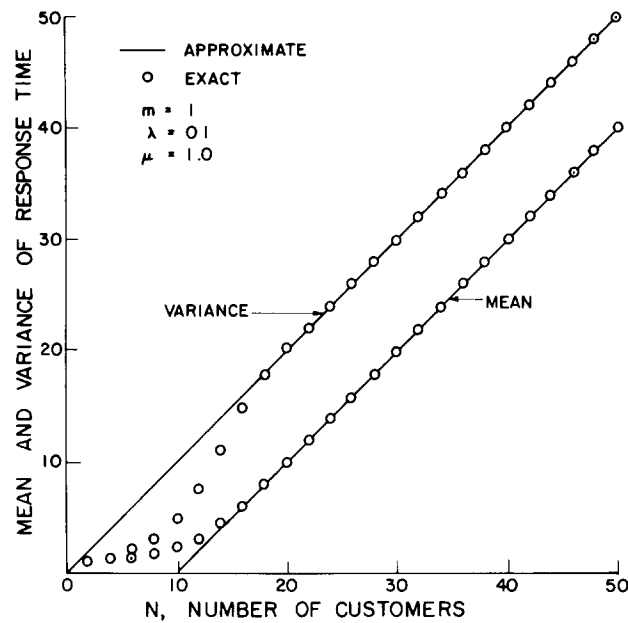


Figure 2. Mean and variance of response time vs. N .

of this distribution are given by:

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

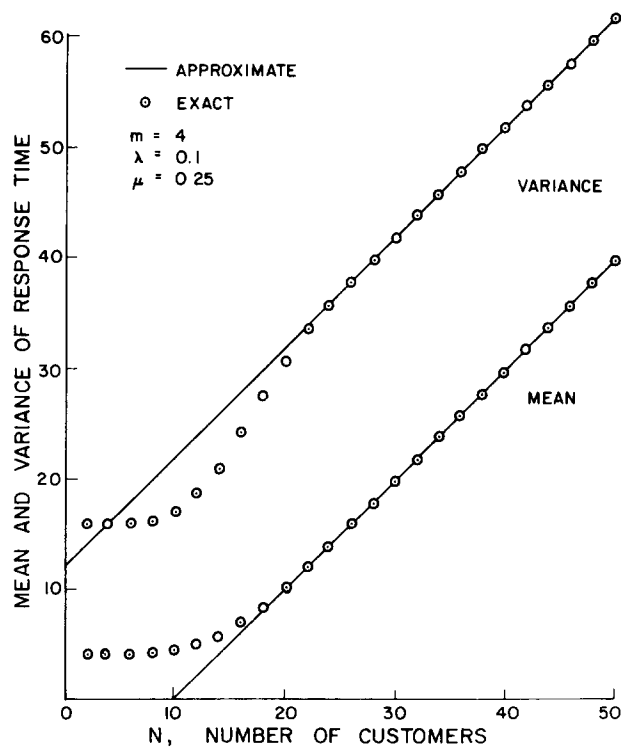
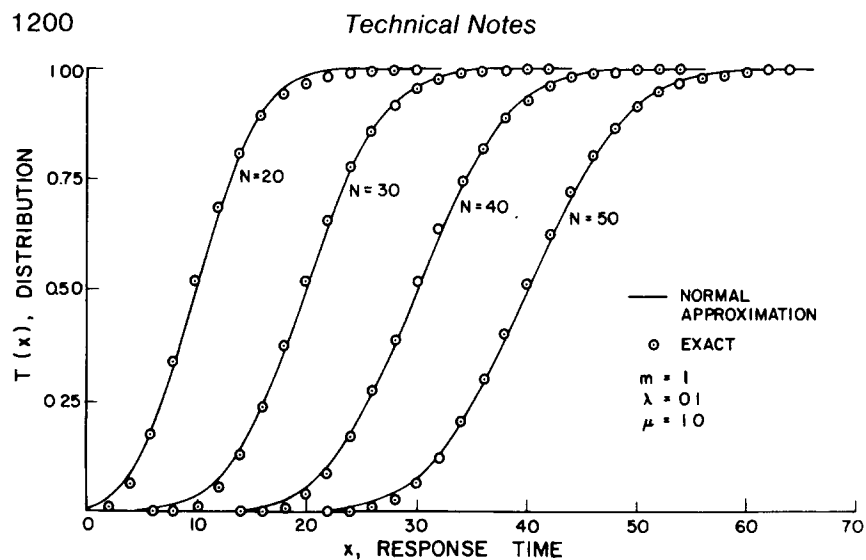
$$\sigma_T^2 = 1/\mu^2 + (N - m)/(m\mu)^2. \quad (10)$$

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

3. NUMERICAL EXAMPLES

In this section, numerical examples are presented to compare our normal approximation with the exact analysis of Sekino [6, 7]. In our



first example, the parameter values are: $m = 1$, $\lambda = 0.1$, $\mu = 1.0$. 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: $m = 4$, $\lambda = 0.1$, $\mu =$

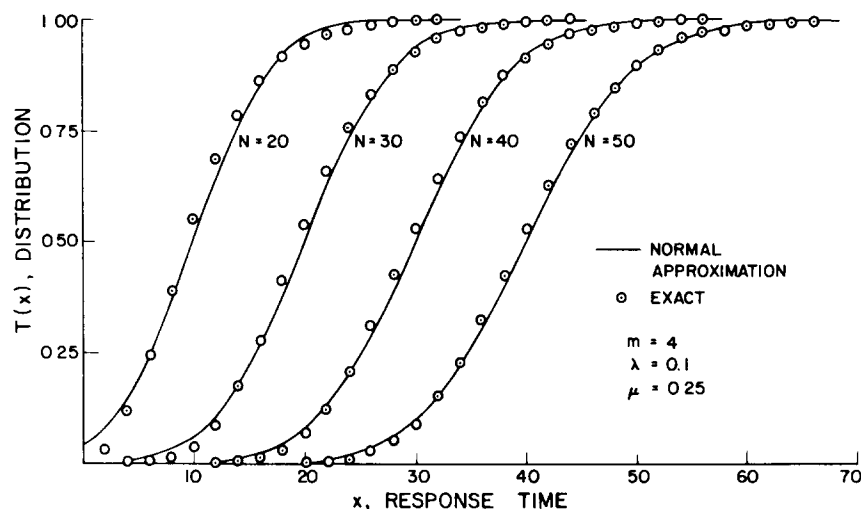


Figure 5. Response time distribution.

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

4. 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 (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 [5-7]. 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 .

ACKNOWLEDGMENT

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

REFERENCES

1. D. IGLEHART, "Limiting Diffusion Approximations for the Many Server Queue and the Repairman Problem," *J. Appl. Probability* **2**, 429-441 (1965).
2. E. LASSETTRE AND A. SCHERR, "Modeling the Performance of the OS/360 Time-Sharing Option (TSO)," *Statistical Computer Performance Evaluation*, pp. 57-72, W. Freiberger (ed.), Academic Press, New York, 1972.
3. A. PAPOULIS, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1965.
4. T. SAATY, *Elements of Queueing Theory*, McGraw-Hill, New York, 1961.
5. A. SCHERR, "An Analysis of Time-Shared Computer Systems," Project MAC, Massachusetts Institute of Technology, MAC-TR-18, June 1965.
6. A. SEKINO, "Performance Evaluation of Multiprogrammed Time-Shared Computer System," Project MAC, Massachusetts Institute of Technology, MAC-TR-103, September 1972.
7. A. SEKINO, "Response Time Distribution of Multiprogrammed Time-Shared Computer Systems," in *Proceedings of the 6th Annual Princeton Conference on Information Sciences and Systems*, pp. 613-619, 1972.