AN APPROXIMATION FOR THE AVERAGE WAITING TIME IN A G/G/c QUEUE

By

Nancy Hubing

CENTER FOR COMMUNICATIONS AND SIGNAL PROCESSING

Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC

June 1984

CCSP-TR-84/5

ABSTRACT

HUBING, NANCY ELLEN. An Approximation for the Average Waiting Time in a G/G/c Queue. (Under the direction of Arne Nilsson.)

It is extremely difficult to solve for the exact average waiting time in a queue with general service and interarrival time distributions. Since many types of distributions occur in computer networks, it is important to have accurate solutions to general queueing systems. Limited tables of exact waiting times in queues with certain complex distributions have been published (3). The tables were derived using lengthy computer programs and excessive amounts of computer time.

Because of the limitations and difficulties involved in using such tables, several approximations have been suggested (11, 13). This work presents a new approximation in which the average waiting time in a queue with general service and interarrival time distributions is computed based on the waiting times in less general queues. Application of this new approximation requires the use of a short computer program and a minimal amount of computer time. The resulting accuracy exceeds the existing approximations in most cases of interest.

TABLE OF CONTENTS

Page

LIST OF TABLES.................................. v

LIST OF FIGURES.................................. vii

LIST OF SYMBOLS.................................. viii

I. INTRODUCTION.................................. 1

II. BACKGROUND AND ASSUMPTIONS.................................. 5

A. The G/G/c Queue.................................. 5
B. Service and Interarrival Time Distributions... 6
1. Exponential Distribution.................................. 6
2. Deterministic Distribution.................................. 8
3. Erlangian Distribution.................................. 9
4. Other Distributions.................................. 11

III. EXISTING APPROXIMATIONS FOR $W_{G/G/c}$.................................. 14

A. Heavy-Traffic Approximation.................................. 14
B. Sakasegawa's Approximation.................................. 16
C. Page's Approximation.................................. 17
D. Diffusion Approximation.................................. 18
E. Comments.................................. 20

IV. THE NEW APPROXIMATION.................................. 22

A. Deriving the Coefficients.................................. 22
B. M/G/c Approximation.................................. 29
C. G/M/c Solution.................................. 31
D. Results.................................. 32

V. COMPARISONS.................................. 35

VI. CONCLUSION.................................. 49

VII. REFERENCES.................................. 51

VIII. APPENDICES.................................. 53

Appendix I. Notation.................................. 53
Appendix II. A Fortran Program for the New Approximation.................................. 54

LIST OF TABLES

Page

1. Heavy-Traffic Approximation for $E_3/E_2/2$ Queue...... 15
2. Sakasegawa's Approximation for $E_3/E_2/2$ Queue....... 16
3. Page's Approximation for $E_3/E_2/2$ Queue............ 18
4. Bell Lab's Approximation (16) for $E_3/E_2/2$ Queue.... 21
5. Exact Mean Waiting Times (units of service time)... 23
6. Error in Assuming A=-0.1109, B=0.7483 in $E_3/E_2/2$... 24
7. The Coefficients of Eq. 23, A(m,k,c) and B(m,k,c).. 25
8. Error in Applying Eq. 30 to $M/E_k/c$................ 31
9. The New Approximation for $E_3/E_2/2$ Queue........... 33
10. Waiting Time Errors in 2/2/1 Queue for New, Page, and Sakasegawa Approximations.............. 36
11. Waiting Time Errors in 4/2/1 Queue................ 36
12. Waiting Time Errors in 16/2/1 Queue............... 36
13. Waiting Time Errors in 2/2/2 Queue................ 37
14. Waiting Time Errors in 9/2/2 Queue................ 37
15. Waiting Time Errors in 2/2/3 Queue................ 37
16. Waiting Time Errors in 3/2/3 Queue................ 38
17. Waiting Time Errors in 2/2/4 Queue................ 38
18. Waiting Time Errors in 9/2/4 Queue................ 38
19. Waiting Time Errors in 2/2/5 Queue................ 39
20. Waiting Time Errors in 9/2/5 Queue................ 39
21. Waiting Time Errors in 2/2/8 Queue................ 39
22. Waiting Time Errors in 3/2/8 Queue................ 40
23. Waiting Time Errors in 2/3/1 Queue................ 40

Page

24. Waiting Time Errors in 16/3/1 Queue................ 40

25. Waiting Time Errors in 2/3/2 Queue................ 41

26. Waiting Time Errors in 4/3/2 Queue................ 41

27. Waiting Time Errors in 2/4/1 Queue................ 41

28. Waiting Time Errors in 4/4/1 Queue................ 42

29. Waiting Time Errors in 9/2/1 Queue................ 42

30. Waiting Time Errors in 4/3/1 Queue................ 42

LIST OF FIGURES

Page

1. The G/G/c Queue 5

2. Exponential Distribution (pdf) 7

3. K-Stage Erlangian Service Time Distribution Representation 10

4. K-Stage Hyperexponential Service Time Distribution Representation 13

5. c*B vs. c Holding m and k Constant 27

6. % Error vs. ρ for 2/2/1 Queue for New, Page, and Sakasègawa Approximations 44

7. Log(% Error) vs. ρ for 2/2/4 Queue 44

8. Log(% Error) vs. ρ for 2/2/8 Queue 45

9. Log(% Error) vs. ρ for 9/2/2 Queue 45

10. Log(% Error) vs. ρ for 16/3/1 Queue 46

11. Log(% Error) vs. ρ for 4/3/2 Queue 46

12. Log(% Error) vs. ρ for 4/4/1 Queue 47

LIST OF SYMBOLS

c = number of parallel servers in a queueing system

c_x^2 = squared coefficient of variation

k = number of Erlangian stages in service time process

λ = rate of arrival of customers

m = number of Erlangian stages in interarrival time process

μ = rate of service of customers

ρ = server utilization

σ_x^2 = variance of the random variable x

t = time

W = average waiting time of a customer in a queue

AN APPROXIMATION FOR THE AVERAGE WAITING TIME IN A G/G/c QUEUE

I. INTRODUCTION

Standing in line at a grocery checkout or a bank teller window are examples of the common experience of waiting in line for a service. In a similar manner messages are queued in a computer network until a particular service can be rendered. The queueing delay encountered by a message is often used as a measure of the performance of the network. It is important to have an accurate value for the average time a message spends waiting at each queue in the network so that network performance can be evaluated.

The amount of time a message spends waiting in a queue is dependent upon the service time distribution and the message interarrival time distribution. When these two distributions are complex, it can be difficult to solve for the exact average waiting time in the queue. Hillier and Yu have published tables (3) of exact values of queue length for systems with Erlangian service and interarrival time distributions. These tables can be used to find queue waiting times for a range of values for the server utilization and the parameters of the system. Exact values not given in the tables can be interpolated if the system falls within the given ranges. If the system is not within the given ranges, the tables cannot be used to obtain accurate values of waiting time. This limitation of the tables prevents their use in some instances.

Another difficulty encountered in using tables of exact waiting times is that they cannot be easily incorporated into a computer program for solving queueing networks. In evaluating and comparing computer networks, it is desirable to have a simple and accurate approximation for waiting time that can be incorporated into such a program. Because of the limitations inherent in using tables of exact values, several approximations for average queueing times have been suggested and evaluated (11, 13).

The existing approximations for average waiting time in queues with Erlangian distributions perform very well for values of server utilization greater than 90%. However, practical networks commonly operate with values of server utilization between 40% and 90%. The waiting time errors produced by the existing approximations increase rapidly as server utilization decreases. The new approximation presented in this work produces smaller errors for utilizations less than 90%, while maintaining the accuracy of the existing approximations for high utilizations.

The new approximation is based in the often-used concept of combining the waiting time solutions of less general queues to approximate the waiting time in a general queue. The two less general systems that will be used are the G/M/c and M/G/c queues. (An explanation of this notation may be found in Appendix I). The approximation is of the form

$$W_{G/G/c} = A*W_{M/G/c} + B*W_{G/M/c} \qquad (1)$$

where W() represents the average waiting time in the indicated queueing system. A and B are functions of the parameters of the G/G/c queue. In this work examples will be limited to systems with Erlangian distributions (Em/Ek/c systems).

In section II some basic queueing theory will be discussed including common service and interarrival time distributions for queueing systems. Several existing approximations are described in section III, including the Heavy-Traffic approximation, the Diffusion Approximation, Page's Approximation (11), and Sakasegawa's Approximation (13). The accuracy and ease of application of each approximation is discussed.

Section IV contains the derivation of the new approximation. The tables of exact values by Hillier and Yu (3) are used to derive and check the approximation. In

Section V comparisons are made between the new and existing approximations based on accuracy and ease of application.

The conclusion summarizes results and presents possible areas of future research for this type of approximation. Appendix I explains notation and Appendix II contains a listing of a Fortran program that accomplishes the new approximation.

II. BACKGROUND AND ASSUMPTIONS

A. The G/G/c Queue

The queueing system this work examines is modeled as shown in fig. 1. The queue is of infinite length and is served on a first-come, first-served basis by the c identical servers. If any of the c servers are idle then the customer at the head of the queue (if any) begins service immediately.

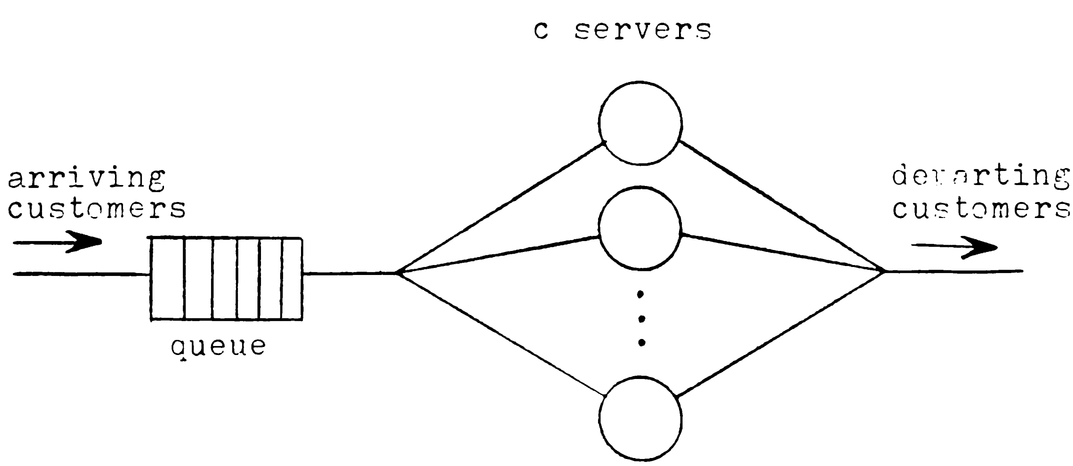

Fig. 1. The G/G/c Queue

Customers arrive according to a general distribution of the time between arrivals (interarrival time). This general distribution can be of any form, and several of the more useful forms will be discussed. The service time for a customer by one of the c servers is similarly described by a general distribution. A customer departs the system as soon as service is completed.

The exact solution for average waiting time in the queue of fig. 1 is not known. Exact solutions can be found for certain less general distributions such as those that will be described next.

B. Service and Interarrival Time Distributions

1. Exponential Distribution

The arrival of customers to a queue is described by a Poisson process, if the probability of k arrivals in any interval of length t can be expressed as follows;

$$P_k(t)=(\lambda t)^k e^{-\lambda t}/k! \quad k\geq 0, t\geq 0 \qquad (2)$$

where λ is the customer arrival rate. For a Poisson arrival process the time between adjacent arrivals is exponentially distributed. The probability density function (pdf) for an exponentially distributed random variable is shown in fig. 2.

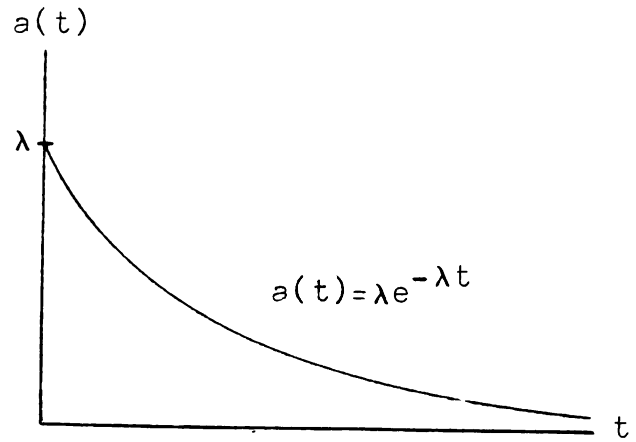

Fig. 2. Exponential Distribution (pdf)

When describing distributions of random variables, it is helpful to define the squared coefficient of variation

$$c_x^2=\sigma_x^2/m^2 \qquad (3)$$

where σ_x is the standard deviation and m is the mean of the random variable. For the important case of the exponentially distributed random variable, $c_x^2=1$. For other distributions the value of the coefficient of variation indicates important properties of the random variable.

The exponential distribution has the property that the past history of the random variable does not affect predictions of future behavior. This "memoryless" property greatly simplifies calculations involving exponential distributions.

The service time in a queueing system can also be exponentially distributed with a pdf written

$$b(t)=\mu e^{-\mu t} \quad t\geq 0 \tag{4}$$

where μ is the service rate. When both the service time and interarrival time are exponentially distributed (M/M/c queue), the queueing system can be solved for the exact average waiting time (8, 5). A queueing system with interarrival time described by a general distribution and exponentially distributed service times (G/M/C) can also be solved exactly (14), if the Laplace transform of the distribution function of the arrival process can be found. However, if the service time is a general distribution and the interarrival time is an exponential distribution (M/G/c), no straightforward solution for average waiting time is known. In this case various accurate approximations for average waiting time have been found (12).

2. Deterministic Distribution

If the interarrival time in a queueing system is deterministic, then the time between arrivals is constant. Similarly a deterministic service time means the service time is constant. For a constantly distributed random variable the variance is zero so the squared coefficient of variation is zero. A queue with constant service time and exponentially distributed interarrival times (M/D/c), can be solved for the exact average waiting time (12).

3. Erlangian Distribution

The k-stage Erlangian service time distribution has a pdf of the form

$$f_s(t)=k\mu(k\mu t)^{k-1}e^{-k\mu t}/(k-1)! \quad t\geq 0 \tag{5}$$

where μ is the service rate and k is a positive integer. This type of distribution is commonly observed in actual queueing networks. Notice that for k=1 the distribution reduces to

$$f_s(t)=\mu e^{-\mu t} \quad t\geq 0 \tag{6}$$

which is the same as the exponential service time distribution of eq. 4.

The k-stage Erlangian distribution is often represented by exponential servers as shown in fig. 3. The server consists of k exponential servers in series, each with a service rate of $k\mu$. Only one customer at a time can be inside the box, so the service rate is μ.

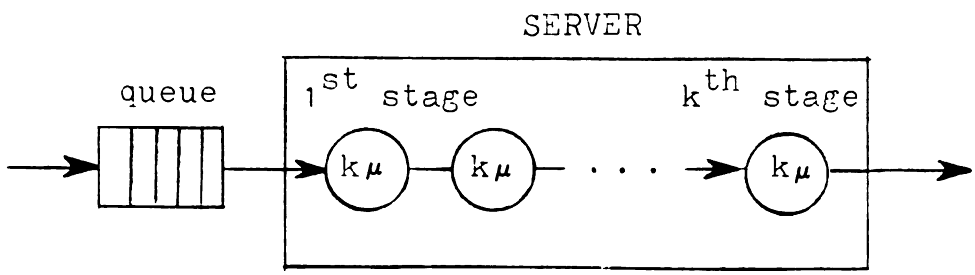

Fig. 3. K-stage Erlangian Service Time Distribution Representation

As k approaches infinity, the time spent in service approaches the constant $1/\mu$. Thus the Erlangian distribution can be used to represent both the exponential and the deterministic distributions by adjusting the number of stages, k. Many other useful distributions can be described for the integer values of $k>1$.

The m-stage Erlangian interarrival time distribution is of the form

$$f_a(t)=m\lambda(m\lambda t)^{m-1}e^{-m\lambda t}/(m-1)! \qquad t\geq 0 \qquad (7)$$

which is analagous to eq. 5. For many networks, a model having Erlangian interarrival time and service time distributions ($E_m/E_k/c$) can describe all queues of interest in the network. For this reason the $E_m/E_k/c$ queue will be used to demonstrate examples in this work.

The variance of a random variable described by an Erlangian distribution is

$$\sigma_x^2 = m^2/r \qquad (8)$$

where m is the mean and r is the number of stages. Thus the squared coefficient of variation is 1/r. Since r is a positive integer, c_x^2 is an element of the set {1, 1/2, 1/3, 1/4, . . . } for the Erlangian distribution.

4. Other Distributions

The Gamma distribution is a generalization of the Erlangian distribution with a pdf of the form

$$f(t)=\alpha(\alpha t)^{\beta-1}e^{-\alpha t}/\Gamma(\beta) \quad t\geq 0 \qquad (9)$$

where α and β are both real and positive and $\Gamma(\beta)$ is the gamma function

$$\Gamma(\beta)=\int_0^\infty y^{\beta-1}e^{-y}dy \ . \qquad (10)$$

When β is a positive integer eq. 9 reduces to the Erlangian distribution of eq. 5.

The squared coefficient of variation for the gamma distribution is $1/\beta$. Since β can be any real positive number, the coefficient of variation can also be any real positive number with this distribution. Therefore the gamma distribution can describe random variables for any $c_x^2>0$, whereas the Erlangian distribution can only describe random variables with $0<c_x^2\leq 1$. The gamma distribution is more general and more complex than the Erlangian distribution, but in many practical cases the Erlangian distribution is sufficient to accurately describe the random variable.

Sometimes the distribution of a random variable is best described by a mixture of exponential distributions. The hyperexponential service time distribution is of the form

$$f(x)=\sum_{i=1}^{k} \pi_i \mu_i e^{-\mu_i x} \qquad (11)$$

where μ_i is the service rate of the i^{th} exponential server and π_i is the probability that a customer will visit that server. Fig. 4 shows the exponential server representation of the hyperexponential distribution. One customer at a time enters the box and is served by the i^{th} server with probability π_i before departing the box.

The squared coefficient of variation for the hyperexponential distribution is always greater than or equal to one, and is a function of the service rates and probabilities π_i (8). This distribution arises in practice

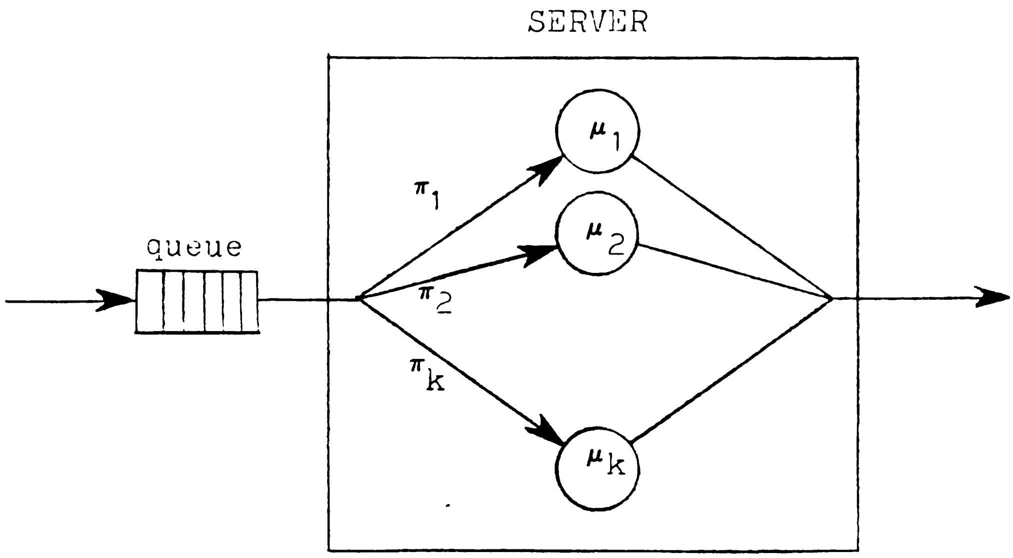

Fig. 4. K-Stage Hyperexponential Service Time Distribution Representation

when there are different types of customers requiring varying lengths of service time. The gamma and hyperexponential distributions will not be covered by the examples in this work.

III. EXISTING APPROXIMATIONS FOR $W_{G/G/c}$

A. Heavy-Traffic Approximation

For the G/G/c queue, the server utilization is defined as

$$\rho=\lambda/(c\mu) \qquad (12)$$

where λ is the arrival rate and μ is the service rate of one of the parallel servers. If ρ approaches one (but remains strictly less than one to preserve stability), the system is said to be in a heavy-traffic state. For the single server (G/G/1) heavy-traffic case it has been shown (6) that

$$W=\lambda(\sigma_a^2+\sigma_s^2)/(2(1-\rho)) \qquad (13)$$

where σ_a^2 is the variance of the interarrival time and σ_s^2 is the variance of the service time. It has also been shown (6) that the upper bound on the mean waiting time in any G/G/1 queue is the right hand side of eq. 13.

For the G/G/c heavy-traffic case it has been established (9) that

$$W=\lambda(\sigma_a^2+\sigma_s^2/c^2)/(2(1-\rho)) \ . \qquad (14)$$

Since so little is known about the G/G/c queue, this approximation is useful for worst-case network analysis and comparison of approximations. It has been suggested (4) but not proven that eq. 14 is an upper bound on the mean waiting time in any G/G/c queue.

Using eq. 14, eq. 11, and the definition of coefficient of variation we obtain

$$W=\rho c(c_a^2/(\rho c)^2+c_s^2/c^2)/((2(1-\rho)) \qquad (15)$$

for the heavy-traffic assumption. We also know that for a $E_m/E_k/c$ queue $c_a^2=1/m$ and $c_s^2=1/k$. Table 1 contains exact mean waiting times for various values of ρ in the $E_3/E_2/2$ queue, and the Heavy-Traffic approximation (eq. 15) for this queue. Obviously this simple approximation can only be applied when ρ is very close to one. Unfortunately this is rarely the case in practical networks.

ρ	Exact Average Waiting Time(3)	Heavy-Traffic Approximation	% Error
0.3	0.01180	0.45040	3716.9
0.5	0.07914	0.45833	479.1
0.7	0.31361	0.68849	119.5
0.9	1.6670	2.0509	23.0
0.95	3.7425	4.1294	12.1

Table 1. Heavy-Traffic Approximation for $E_3/E_2/2$ Queue

B. Sakasegawa's Approximation

A simple approximation for the mean waiting in the $E_m/E_k/c$ queue is given by Sakasegawa (13),

$$W_{Em/Ek/c}=(1/m+1/k)\rho^{P}/(2c\rho(1-\rho)) \qquad (16)$$

where $p=(2(c+1))^{0.5}$. This approximation overestimates the average waiting time but improves in accuracy as ρ increases. Table 2 shows the average waiting time in the $E_3/E_2/2$ queue as given by Sakasegawa's approximation. This approximation produces better estimates for the average waiting time than the Heavy-Traffic Approximation (table 1). Sakasegawa's Approximation is very easy to apply and can be used in rough analysis when high accuracy is not important.

ρ	Exact Average Waiting Time(3)	Sakasegawa's Approximation	% Error
0.3	0.01180	0.05197	340.4
0.5	0.07914	0.15256	92.8
0.7	0.31361	0.41410	32.0
0.9	1.6670	1.7883	7.28
0.95	3.7425	3.8681	3.36

Table 2. Sakasegawa's Approximation for $E_3/E_2/2$ Queue

C. Page's Approximation

An interesting approximation for the average waiting time in a $E_m/E_k/c$ queue was derived by Page (11) and involves combinations of the waiting times in less general queues.

$$W_{Em/Ek/c}=(1-1/m)(1/k)W_{D/M/c}+ \\ +(1/m)(1-1/k)W_{M/D/c}+(1/m)(1/k)W_{M/M/c} \qquad (17)$$

To obtain this approximation the various combinations of the limiting values of m and k were examined as follows:

$$m=1,\ k=1 \implies W_{Em/Ek/c}=W_{M/M/c} \qquad (18a)$$

$$m=1,\ k\rightarrow\infty \implies W_{Em/Ek/c}=W_{M/D/c} \qquad (18b)$$

$$m\rightarrow\infty,\ k=1 \implies W_{Em/Ek/c}=W_{D/M/c} \qquad (18c)$$

$$m\rightarrow\infty,\ k\rightarrow\infty \implies W_{Em/Ek/c}=W_{D/D/c}=0 \qquad (18d)$$

Eq. 17 is exact for all of the limiting cases of eqs. 18.

The three queues D/M/c, M/D/c, and M/M/c can all be solved for the exact mean waiting time (12). The solutions for the D/M/c and M/D/c queues are rather complex, so it might be easier to use tables of exact values for these queues when applying eq. 17. Page's approximation has been tested for values of m and k other than the limiting cases. Table 3 compares exact values of mean waiting time with Page's Approximation for the $E_3/E_2/2$ queue. Exact values were used for $W_{D/M/c}$, $W_{M/D/c}$, and $W_{M/M/c}$ (11) when applying eq. 17.

The approximation tends to overestimate the average waiting time but improves for larger values of ρ just as in the Heavy-Traffic and Sakasegawa Approximations. However, the accuracy of Page's Approximation is much higher. Page's Approximation (eq. 17) is one of the most accurate available in the literature for estimating the mean waiting time in multi-server queues. The only drawback is that it is rather difficult to apply.

D. Diffusion Approximation

Another approach to solving for the average waiting time is to approximate the stochastic processes themselves (the arrival and departure processes). Basically the diffusion approximation replaces the discrete queueing process by a continuous one in such a way that the important

ρ	Exact Average Waiting Time(3)	Page's Approximation	% Error
0.3	0.01180	0.02729	131.3
0.5	0.07914	0.10665	34.8
0.7	0.31361	0.35095	11.9
0.9	1.6670	1.7121	2.71
0.95	3.7425	3.7856	1.15

Table 3. Page's Approximation for $E_3/E_2/2$ Queue

characteristics of the discrete process are retained (2).

Let N(t) be the number of customers in the queueing system at time t, and $P_n(t)$ be the probability that N(t)=n at time t. The diffusion approximation involves replacing the discrete valued variable N(t) with the continuous variable X(t). The density function of X(t), f(x,t) = $P\{x \le X(t) \le x+dx\}$, will be used to recover estimates of $P_n(t)$.

If X(t) behaves as a diffusion process, then f(x) obeys the steady state Fokker-Planck equation

$$1/2 \; d^2/dx^2 \; \{\sigma^2(x)f(x)\} - d/dx \; \{m(x)f(x)\} = 0 \qquad (19)$$

where m(x) and $\sigma^2(x)$ are the infinitessimal mean and variance of the diffusion process (10).

To apply the diffusion equation (eq. 19) to multi-server queueing systems involves the determination of appropriate choices for m(x) and $\sigma^2(x)$. One possibility (2) is

$$m(x) = \lambda - \min(x,c)\mu \qquad x \ge 0$$
$$\sigma^2(x) = \lambda^3\sigma_a^2 + \min(x,c)\mu^3\sigma_s^2 \qquad x \ge 0 \qquad (20)$$

where $1/\lambda$ and $1/\mu$ are the mean interarrival time and mean service time and σ_a^2 and σ_s^2 are the variance of the interarrival time and the variance of the service time respectively.

The solution of eq. 19 involves the application of several boundary conditions and will not be discussed in detail (see 1, 2, 10). P_n can be recovered by the discretization process

$$P_n=\int_{n-.5}^{n+.5} f(x)dx \qquad (n=1,2,\ldots)$$

after f(x) is known (7).

Using the procedure for solving for f(x) in (2) yields results for the $E_m/E_k/c$ queue which are usually as good or better than Sakasegawa's Approximation. By using an additional condition not used in (2), results were obtained in (10) which were almost as good as Page's Approximation.

Like all of the approximations discussed, the diffusion approximation gives very good results for ρ close to one but decreases in accuracy for decreasing ρ. Current and future research on the diffusion approximation for multi-server queues is expected to produce even more accurate results.

E. Comments

Most simple approximations for the mean waiting time in the $E_m/E_k/c$ queue are based on heavy-traffic limit theorems and yield results with fairly high errors. One such approximation is of the form

$$W_{Em/Ek/c}=(1/k+1/m)*W_{M/M/c}/2 \qquad (21)$$

and is currently used in a software package developed at Bell Labs for their Queueing Network Analyzer (16). The performance of eq. 21 for the $E_3/E_2/2$ queue is shown in table 4. This approximation performs slightly better than Sakasegawa's Approximation (table 2) for the $E_3/E_2/2$ system. The accuracy of the Bell Lab's Approximation decreases rapidly with increasing values of k or m. Sakasegawa's Approximation outperforms the Bell Labs Approximation in these cases.

Looking back at tables 1, 2, and 3 gives a good idea of the general performance of the Heavy-Traffic, Sakasegawa, and Page approximations for $E_m/E_k/c$ queues. These three approximations will be used in the following sections to evaluate the accuracy of the new approximation for various $E_m/E_k/c$ queues.

ρ	$W_{M/M/2}$	(exact) $W_{E3/E2/2}$	(approx.) $W_{E3/E2/2}$	% Error
0.3	0.0989	0.0118	0.0412	249.2
0.5	0.3333	0.0791	0.1389	75.5
0.7	0.9608	0.3136	0.4003	27.6
0.9	4.2632	1.6670	1.7763	6.56

Table 4. Bell Lab's Approximation (16) for $E_3/E_2/2$ Queue

IV. THE NEW APPROXIMATION

A. Deriving the Coefficients

The new approximation for the G/G/c queue presented in this work is of the form

$$W_{G/G/c} = A*W_{M/G/c} + B*W_{G/M/c} \qquad (22)$$

where A and B are functions of the parameters of the G/G/c queue. Erlangian distributions will be used for the interarrival and service time distributions to determine A and B so eq. 22 becomes

$$W_{Em/Ek/c} = A*W_{M/Ek/c} + B*W_{Em/M/c} \qquad (23)$$

where A and B are functions of m, k, and c.

Tables of exact values of mean queue lengths for $E_m/E_k/c$ queues have been published (3). The mean waiting time can easily be obtained from the mean queue length using Little's formula

$$W = L/\rho c. \qquad (24)$$

Using these tables A and B can be determined for various values of m, k, and c.

As an example let us examine the $E_3/E_2/2$ queue (m=3, k=2, c=2). From the tables (3) we can obtain exact values for $W_{E3/E2/2}$, $W_{M/E2/2}$, and $W_{E3/M/2}$ for a range of values of ρ, the server utilization. In order to solve for A and B for the $E_3/E_2/2$ queue, choose two reasonable values of ρ, such as $\rho=0.6$ and $\rho=0.9$. The exact mean waiting times for these values of ρ for all three queues of eq. 23 are listed in Table 5.

From table 5 and eq. 23 we obtain two equations in the two unknowns, A and B

$$0.16059 = A*0.42876 + B*0.27814$$

$$1.6670 = A*3.2073 + B*2.7030 .$$

Solving for A and B gives

$$A=-0.1109$$

$$B=0.7483 \qquad (25)$$

for m=3, k=2, and c=2.

ρ	$W_{E3/E2/2}$	$W_{M/E2/2}$	$W_{E3/M/2}$
0.6	0.16059	0.42876	0.27814
0.9	1.6670	3.2073	2.7030

Table 5. Exact Mean Waiting Times
(units of service time)

Next we must check to be certain that A and B do not vary too much with ρ for the $E_3/E_2/2$ system. To do this we use the calculated values of A and B in eq. 23 for other values of ρ. Table 6 lists the error in assuming these values of A and B for a range of values of ρ in the $E_3/E_2/2$ system.

By examining table 6 it appears that given A, B, and the exact values for $W_{M/E2/2}$ and $W_{E3/M/2}$, we can accurately approximate $W_{E3/E2/2}$. For $\rho>0.5$ the error in this approximation is less than 1%. For $\rho<0.3$ the error is larger, but these small values of server utilization are rarely of concern in practical multi-server systems. The small waiting times encountered when $\rho<0.3$ add little to the overall delay of a customer.

The next step in developing the new approximation is to repeat the calculations of A and B for other $E_m/E_k/c$ queues.

ρ	$W_{M/E2/2}$	$W_{E3/M/2}$	(Approx eq. 23) $W_{E3/E2/2}$	(Exact) $W_{E3/E2/2}$	% Error
0.3	0.07709	0.02466	0.00991	0.01180	16.0
0.4	0.14715	0.06603	0.03310	0.03460	4.3
0.5	0.25564	0.14257	0.07834	0.07914	1.0
0.7	0.72857	0.52808	0.31438	0.31361	0.25
0.95	6.9526	6.0258	3.7381	3.7425	0.12

Table 6. Error in Assuming A=-0.1109, B=0.7483 in $E_3/E_2/2$

We wish to obtain A(m, k, c) and B(m, k, c) for a wide range of values of m, k, and c. Once this is done we will attempt to formulate equations for A and B in terms of m, k, and c. Using these equations and exact or approximate values for $W_{M/Ek/c}$ and $W_{Em/M/c}$, we will be able to approximate $W_{Em/Ek/c}$.

Using eqs. 23 and 24 and the tables of exact mean waiting times (3) table 7 was compiled. For all entries ρ=0.6 and ρ=0.9 were the values used to solve for A and B in eq. 23. For other values of ρ errors similar to those shown in table 6 were obtained. In general, for the larger values of c (the number of servers) the error in assuming A and B

m	k	c	A(m, k, c)	B(m, k, c)	m	k	c	A(m, k, c)	B(m, k, c)
2	2	1	-0.1709	0.8367	2	2	5	-0.0622	0.72945
3	2	1	-0.1676	0.8120	3	2	5	-0.0782	0.71196
4	2	1	-0.1650	0.7958	4	2	5	-0.0789	0.69215
16	2	1	-0.1560	0.7447	9	2	5	-0.0722	0.64102
2	2	2	-0.0966	0.7629	2	3	1	-0.2348	0.76308
3	2	2	-0.1109	0.7483	3	3	1	-0.2258	0.72356
4	2	2	-0.1123	0.7326	4	3	1	-0.2191	0.69724
9	2	2	-0.1087	0.6923	9	3	1	-0.2033	0.63834
2	2	3	-0.0750	0.7415	16	3	1	-0.1953	0.61058
3	2	3	-0.0924	0.7277	2	3	2	-0.1264	0.66726
2	2	4	-0.0660	0.7328	3	3	2	-0.1438	0.64199
3	2	4	-0.0834	0.7178	4	3	2	-0.1435	0.61696
4	2	4	-0.0848	0.6994	2	4	1	-0.2666	0.72085
9	2	4	-0.0796	0.6518	3	4	1	-0.2527	0.67183
9	2	1	-0.1590	0.7608	4	4	1	-0.2424	0.63884

Table 7. The Coefficients of Eq. 23, A(m, k, c) and B(m, k, c)

were independent of ρ was largest. This might have been expected since the waiting time in those systems is smaller than in a system with fewer servers for a given value of ρ. In practical networks containing $E_m/E_k/c$ queues, we would expect that a system with 5 servers could operate at a higher value of ρ than a system with 1 server. Since the error in applying eq. 23 seems to always decrease with increasing ρ, we shall not be concerned that systems with a larger number of servers have slightly higher errors.

Now table 7 must be used to attempt to derive equations for the coefficients A and B. The equations we derive will only approximate A and B for each $E_m/E_k/c$ queue. This introduces error which will be reflected in the new approximation.

In an attempt to discover some kind of a pattern in the A and B values of table 7, numerous graphs were plotted. The method used was to plot A (or B) as a function of one of the three variables, while holding the other two variables constant. The resulting graphs were examined for a linear or exponential pattern of some kind. Since so much trial and error was involved, there is no need to go into great detail about the derivation of the equations for A and B. Obviously any equations derived in such a way will not be unique and may not be the best possible equations to describe the data of table 7. A tradeoff must be made between complexity of the equations and the accuracy with which the equations describe the data of table 7.

A brief description of the derivation of the equation for B(m, k, c) will be presented. Numerous plots were examined but one in particular stood out as a good starting point in determining an equation. The graph of B(c) holding m and k constant looks like a 1/c type curve in all cases. To check this, a plot of c*B vs. c was drawn for k=2 as shown in fig. 5. Similar plots were drawn for k=3 and k=4.

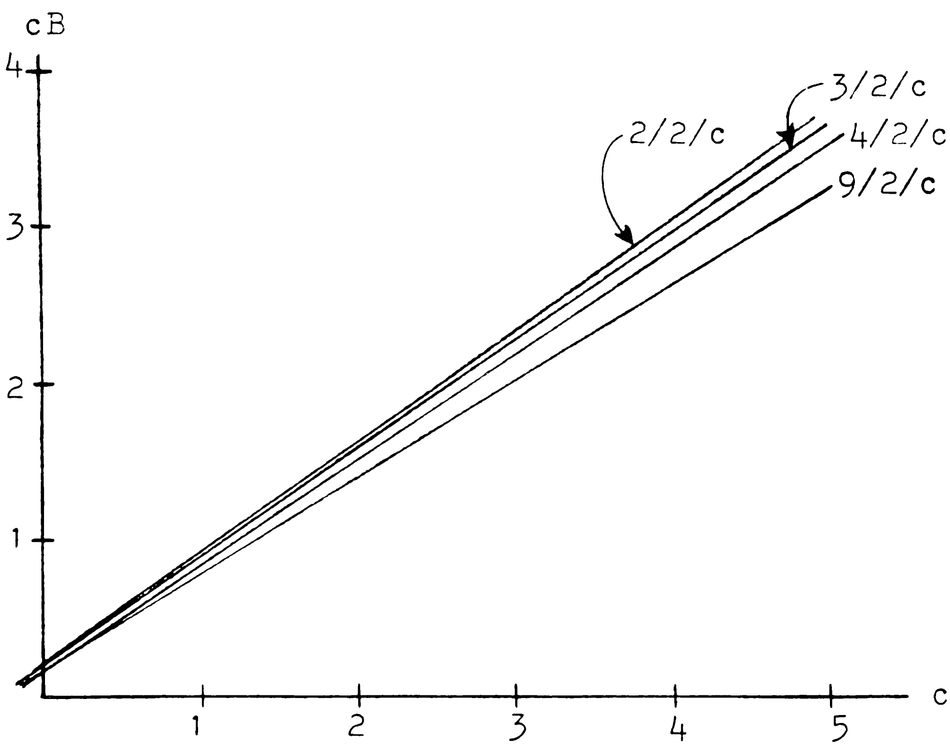

Fig. 5. c*B vs. c Holding m and k Constant

From fig. 5 we can write

$$c*B = f_s(m,k)*c+f_i(m,k)$$

or equivalently

$$B = f_s(m,k)+f_i(m,k)/c \qquad (26)$$

where $f_s(m,k)$ is the slope and $f_i(m,k)$ is the y-intercept of the lines in fig. 5. Now the problem of determining an equation for B is reduced to finding equations for $f_s(m,k)$ and $f_i(m,k)$. Since this involved a similar trial and error approach, no details will be presented. The approximate set of equations found for $f_s(m,k)$ and $f_i(m,k)$ was

$$f_s(m,k)=0.1e^{-(m-1)^2/16}+e^{-k/4}$$

$$f_i(m,k)=0.41e^{-(k-10)^2/50} \qquad (27)$$

With a similar approach for deriving an equation for A, we finally obtain

$$A=-(1-1/k)\exp\{-(1-1/m)[(c/10+1.1)/(m-1)+3-2/c]\} \qquad (28)$$

$$B=0.1e^{-(m-1)^2/16}+e^{-k/4}+(0.41/c)e^{-(k-10)^2/50} . \qquad (29)$$

Rather than presenting the details of checking eqs. 28 and 29 against table 7, we choose to apply the approximation of eq. 23 and check actual waiting times. Before this is done it is necessary to obtain exact or good approximate solutions to $W_{Em/M/c}$ and $W_{M/Ek/c}$ so that the application of the approximation will not involve the tables (3). This will be done in the next two sections.

B. M/G/c Approximation

It is difficult to obtain exact results for the M/G/c queue so a good approximation will be used to test the new approximation. Any error in the M/G/c approximation will be reflected in the new approximation by reducing the accuracy.

Page's Approximation (11) for the $E_m/E_k/c$ queue (eq. 17) can be reduced to an very accurate approximation for the $M/E_k/c$ queue by setting m=1

$$W_{M/Ek/c} = (1-1/k)W_{M/D/c} + (1/k)W_{M/M/c}. \qquad (30)$$

Since the new approximation will only be tested for $E_m/E_k/c$ queues, this approximation for $W_{M/Ek/c}$ could be used if it is highly accurate. We must keep in mind that the error in the new approximation will be increased by any error in the approximations for A, B, $W_{M/Ek/c}$, and $W_{Em/M/c}$. If $W_{M/D/c}$ and $W_{M/M/c}$ can be accurately obtained, it would be straightforward to incorporate eq. 30 into a program for testing the new approximation.

The average queue length in the M/M/c queue can be obtained (12) by first writing down the probabilities of there being any number of customers in the system. The probability of an idle system, P(0), is calculated by summing all of the probabilities and setting the sum equal to 1. The mean queue length can be written

$$L_q = P(0)(\rho^{c+1}/c)/(c!(1-\rho/c)^2) \quad . \qquad (31)$$

We know $W=L_q/\rho c$ so once P(0) is known, eq. 31 can be used to obtain $W_{M/M/c}$.

The exact mean waiting time in the M/D/c queue is more difficult to obtain and a lengthy derivation can be found in (12). Since eq. 30 is an approximation, there is no need to have a more accurate value for $W_{M/D/c}$. We will use the approximation

$$W_{M/D/c} = W_{M/M/c}\{1+(1-\rho)(c-1)[(4+5c)^{.5}-2]/(16\rho c)\}/2 \quad .$$

Now that we have equations for $W_{M/M/c}$ and $W_{M/D/c}$, we can test eq. 30 for accuracy. Table 8 shows that eq. 30 is indeed highly accurate for a range of values of ρ, k, and c. This appears to be the most accurate approximation for $W_{M/Ek/c}$ available in the literature so it will be used for testing eq. 23. The tables of exact values (3) would yield better results, but could not be easily incorporated into a simple computer program, so they will not be used.

ρ	k	c	Exact $W_{M/Ek/c}$	Approx. $W_{M/Ek/c}$	% Error
0.4	2	2	0.1472	0.1467	0.34
0.7	2	2	0.7286	0.7262	0.33
0.9	2	2	3.2073	3.2038	0.11
0.3	4	1	0.2679	0.2679	0.0
0.7	4	1	1.4583	1.4583	0.0
0.9	4	1	5.625	5.625	0.0
0.3	2	5	0.00476	0.00488	2.5
0.7	2	5	0.1947	0.1935	0.62
0.9	2	5	1.1533	1.1509	0.21

Table 8. Error in Applying Eq. 30 to $M/E_k/c$

C. G/M/c Solution

The mean waiting time in the $E_m/M/c$ queue must also be obtained before eq. 23 can be applied. Takacs gives the exact solution for this system (14) in a form that can easily be incorporated into a computer program. If we let

$$\Phi(s)=\int_0^\infty e^{-sx}dF(x)$$

where F(x) is the probability distribution function of the interarrival time process, then the average waiting time in the $E_m/M/c$ queue is

$$W = R/(c(1-\sigma)^2) \qquad (32)$$

where

$$R=[1/(1-\sigma)+\sum_{j=1}^{c}\binom{c}{j}(c(1-\Phi_j)-j)/(c_j(1-\Phi_j)(c(1-\sigma)-j))]^{-1} \qquad (33)$$

$$c_j = \prod_{i=1}^{j} (\Phi_i/(1-\Phi_i)) \tag{34}$$

$$\Phi_j = (cm\rho/(j+cm\rho))^m \tag{35}$$

$$\Phi_v = \Phi(v\mu) \ . \tag{36}$$

For the $E_m/M/c$ queue we have (8)

$$\Phi(s) = (m\lambda/(m\lambda+s))^m . \tag{37}$$

σ is the only root of the equation

$$\sigma = \Phi(c\mu(1-\sigma)) . \tag{38}$$

For this (and all waiting time calculations) we assume without loss of generality that $\mu=1$ so that the waiting time will be in units of service time. Eqs. 32-38 give the solution for the average waiting time in the $E_m/M/c$ queue. Using this and the $M/E_k/c$ approximation of the last section, we can now test the new approximation.

D. Results

The new approximation for the mean waiting time in the $E_m/E_k/c$ queue (eq. 23) uses the approximations for A and B (eqs. 28 and 29). Appendix II contains a listing of a Fortran program which takes as inputs m, k, c, and ρ and gives the approximation for $W_{Em/Ek/c}$ as the output. The approximation for $W_{M/Ek/c}$ (eq. 30) and the exact solution for $W_{Em/M/c}$ (eqs. 32-38) are used in the program. The results obtained for the $E_3/E_2/2$ queue are shown in table 9.

Comparing table 9 with the approximations of tables 1-3 for the $E_3/E_2/2$ queue indicates that the new approximation may perform better in general for moderate values of ρ (ρ<0.9). Since server utilization is usually in the range 0.3<ρ<0.90 for practical queueing networks, the new approximation looks promising.

It was found in some queues that for very low values of ρ (ρ<0.3) the new approximation (eq. 23) sometimes gave a negative value for the mean waiting time. However, the magnitude of the approximate $W_{Em/Ek/c}$ still decreased with decreasing ρ. Therefore the absolute value of the right hand side of eq. 23 was used as the approximation in the program of Appendix II.

The next step is to test the new approximation for a wide range of values of m, k, c, and ρ and compare the

ρ	Exact $W_{E3/E2/2}$	New Approx. $W_{E3/E2/2}$	% Error
0.3	0.01180	0.01169	0.9
0.5	0.07914	0.08390	6.0
0.7	0.3136	0.3294	5.0
0.9	1.6670	1.7302	3.8
0.95	3.7425	3.8735	3.5

Table 9. The New Approximation for $E_3/E_2/2$ Queue

results to Sakasegawa's and Page's approximations. The Heavy-Traffic Approximation always performs worse than the others so it will not be included in the comparisons of the next section.

The only way to check the approximations is to compare the results with the exact values found in the tables (3). Since these tables are limited in m, k, and c, our comparisons will also be limited.

V. COMPARISONS

For convenience the new approximation is repeated here.

$$W_{Em/Ek/c} = A*W_{M/Ek/c} + B*W_{Em/M/c} \quad (39)$$

$$A=-(1-1/k)\exp\{-(1-1/m)[(c/10+1.1)/(m-1)+3-2/c]\} \quad (40)$$

$$B=0.1e^{-(m-1)^2/16}+e^{-k/4}+(0.41/c)e^{-(k-10)^2/50} \quad (41)$$

For the comparisons of this section the approximation of eq. 30 was used for the $M/E_k/c$ waiting time. The percent error in Page's Approximation (eq. 17) and Sakasegawa's Approximation (eq. 16) are included in tables 10-30 for comparison.

By examining tables 10-30 we see that the new approximation always performs better than Page's Approximation for $\rho<=0.7$ and often performs better for $\rho<=0.9$. Page's Approximation is seen to be better than Sakasegawa's Approximation in all cases. The worst results for the new approximation are for high values of m, as in tables 12 and 24 for the 16/2/1 and 16/3/1 queues ($E_{16}/E_2/1$ and $E_{16}/E_3/1$).

For higher values of ρ ($\rho>0.9$) the new approximation is not usually much worse than Page's Approximation and is sometimes better. For example, table 16 (3/2/3 queue) shows that for $\rho=0.95$ the error in the new approximation is 3.96%

rho	New App. % Error	Page App. % Error	Saka. App. % Error	Exact Value
.30	6.13	31.00	63.64	.1310
.50	3.03	12.51	28.08	.3904
.70	2 57	5.39	12.28	1.0391
.90	2.67	1.46	3.23	4.3590
.95	2.73	.69	1.54	9.3561

Table 10. Waiting Time Errors in 2/2/1 Queue for New, Page, and Sakasegawa Approximations

rho	New App. % Error	Page App. % Error	Saka. App. % Error	Exact Value
.30	29.41	78.41	196.81	.0541
.50	8.36	26.84	67.59	.2238
.70	5.41	10.79	26.31	.6927
.90	5.22	2.81	6.47	3.1699
.95	5.31	1.31	3.04	6.9151

Table 11. Waiting Time Errors in 4/2/1 Queue

rho	New app. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	9.13	191.62	770.06	.0139
.50	29.47	49.97	152.17	.1115
.70	14.36	18.24	48.67	.4414
.90	12.35	4.50	10.85	2.2834
.95	12.32	2.08	5.00	5.0891

Table 12. Waiting Time Errors in 16/2/1 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	3.99	69.84	165.59	.0235
.50	.87	21.66	55.07	.1181
.70	.59	7.85	20.46	.4125
.90	.69	1.89	4.79	2.0478
.95	.74	.87	2.22	4.5409

Table 13. Waiting Time Errors in 2/2/2 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	64.78	344.97	1474.96	.0024
.50	13.27	60.41	212.78	.0358
.70	9.15	18.99	59.81	.1900
.90	4.65	4.37	12.41	1.1666
.95	3.84	1.99	5.64	2.6851

Table 14. Waiting Time Errors in 9/2/2 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	5.33	120.64	325.44	.0062
.50	1.08	31.51	83.75	.0511
.70	.81	10.16	27.20	.2275
.90	.81	2.26	5.79	1.2994
.95	.82	1.03	2.63	2.9573

Table 15. Waiting Time Errors in 2/2/3 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	3.98	249.42	729.90	.0026
.50	7.82	50.93	141.90	.0323
.70	5.89	15.28	42.19	.1696
.90	4.28	3.31	8.74	1.0534
.95	3.96	1.50	3.96	2.4328

Table 16. Waiting Time Errors in 3/2/3 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	7.18	187.93	588.80	.0019
.50	1.00	42.51	116.95	.0257
.70	.88	12.47	33.61	.1442
.90	.84	2.58	6.56	.9340
.95	.82	1.16	2.92	2.1741

Table 17. Waiting Time Errors in 2/2/4 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	234.11	1808.69	11860.85	.0001
.50	23.43	121.30	519.16	.0055
.70	14.25	27.42	97.61	.0596
.90	6.73	5.31	16.72	.5211
.95	5.61	2.36	7.35	1.2737

Table 18. Waiting Time Errors in 9/2/4 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	9.82	276.79	1030.76	.0007
.50	.65	54.82	156.37	.0141
.70	.78	14.81	40.04	.0988
.90	.76	2.88	7.23	.7193
.95	.73	1.29	3.16	1.7085

Table 19. Waiting Time Errors in 2/2/5 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	884.36	4017.32	33084.43	.0000
.50	24.87	166.98	772.75	.0025
.70	15.80	32.24	118.25	.0388
.90	6.94	5.75	18.56	.3976
.95	5.72	2.54	8.04	.9970

Table 20. Waiting Time Errors in 9/2/5 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	27.54	749.95	5293.35	.0000
.50	1.50	100.95	329.75	.0031
.70	.11	22.17	60.62	.0408
.90	.24	3.75	8.94	.4077
.95	.20	1.62	3.72	1.0205

Table 21. Waiting Time Errors in 2/2/8 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	27.91	2925.52	23888.90	.0000
.50	21.81	208.06	701.13	.0014
.70	23.73	52.15	123.04	.0245
.90	6.72	7.09	15.67	.3200
.95	5.49	3.25	6.79	.8260

Table 22. Waiting Time Errors in 3/2/8 Queue

rho	New App. % Error	Page App % Error	Saka App. % Error	Exact Value
.30	11.21	43.90	71.22	.1043
.50	6.27	18.64	31.42	.3170
.70	5.52	8.16	13.75	.8547
.90	5.68	2.20	3.62	3.6189
.95	5.77	1.04	1.72	7.7826

Table 23. Waiting Time Errors in 2/3/1 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	506.12	502.36	1527.32	.0052
.50	58.22	101.19	226.58	.0606
.70	18.23	33.46	65.57	.2789
.90	13.51	7.75	13.94	1.5633
.95	13.44	3.57	6.38	3.5351

Table 24. Waiting Time Errors in 16/3/1 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	11.09	93.65	180.51	.0185
.50	5.26	30.30	58.98	.0960
.70	4.57	11.31	21.77	.3401
.90	4.60	2.75	5.09	1.7016
.95	4.65	1.28	2.36	3.7789

Table 25. Waiting Time Errors in 2/3/2 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	21.74	308.14	691.68	.0046
.50	12.99	68.59	145.18	.0436
.70	10.09	22.52	45.67	.1990
.90	6.74	5.17	9.95	1.1385
.95	6.02	2.38	4.56	2.5895

Table 26. Waiting Time Errors in 4/3/2 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	16.45	52.39	75.77	.0914
.50	10.07	22.62	33.43	.2810
.70	9.01	9.94	14.63	.7633
.90	9.13	2.67	3.86	3.2497
.95	9.22	1.27	1.83	6.9967

Table 27. Waiting Time Errors in 2/4/1 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	79.83	172.21	288.41	.0276
.50	18.30	57.49	92.62	.1298
.70	9.95	22.47	34.90	.4324
.90	9.17	5.63	8.42	2.0752
.95	9.31	2.63	3.94	4.5701

Table 28. Waiting Time Errors in 4/4/1 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	83.03	147.08	488.59	.0222
.50	24.30	42.01	119.74	.1391
.70	14.28	15.83	41.03	.5056
.90	12.90	3.97	9.45	2.5127
.95	12.78	1.84	4.38	5.5619

Table 29. Waiting Time Errors in 9/2/1 Queue

rho	New App. % Error	Page App. % Error	Saka App. % Error	Exact Value
.30	54.25	130.14	249.72	.0357
.50	12.50	44.32	82.32	.1600
.70	6.93	17.58	31.42	.5178
.90	6.59	4.47	7.64	2.4387
.95	6.74	2.09	3.58	5.3503

Table 30. Waiting Time Errors in 4/3/1 Queue

while the error in Page's approximation is 1.5%. Either percentage would probably be acceptable in queueing network performance evaluation.

However for lower (and more practical) values of ρ the new approximation is more accurate than the others. Referring again to table 16 (3/2/3 queue) shows that for $\rho=0.7$ the error in the new approximation is 5.89% while the error in Page's Appproximation is 15.3%. An error of 15% might not be acceptable in network performance evaluation. Table 16 is typical of many of the results obtained.

Graphs of the percent error vs. server utilization for some of the systems investigated are shown in figs. 6-12. Fig. 6 shows the percent error on an absolute scale for the Page, Sakasegawa, and new approximations for the 2/2/1 queue. Because of the large range of error the rest of the graphs use a log scale.

All of the graphs show that the new approximation has an advantage over Page's and Sakasegawa's approximations for values of $\rho<0.8$. The worst case of the 16/3/1 queue is graphed in fig. 10. In this case the new approximation still performs better in the range $0.3<\rho<0.8$.

The graphs of figs. 6-12 show that the new approximation does not converge to the exact waiting time as ρ

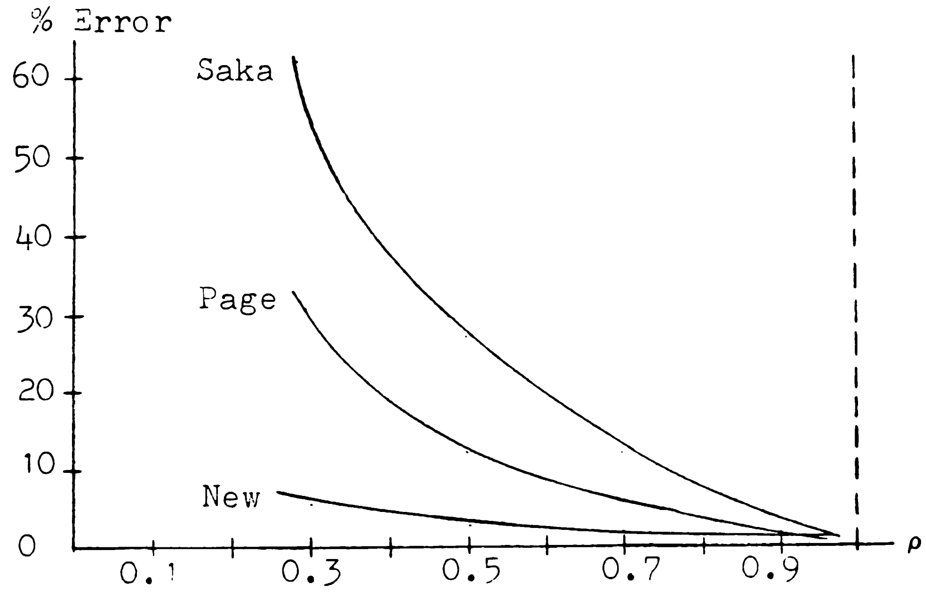

Fig. 6. % Error vs. ρ for 2/2/1 Queue for New, Page, and Sakasegawa Approximations

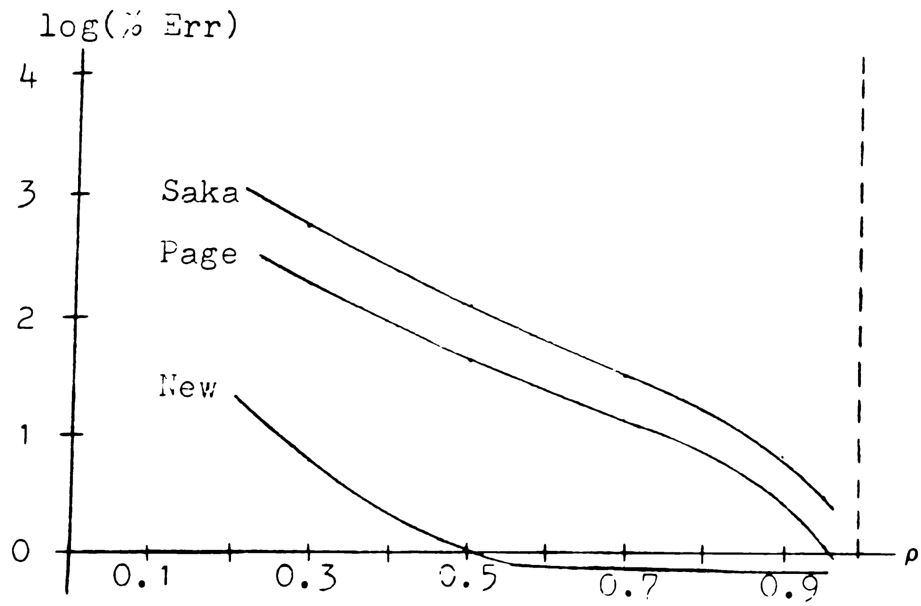

Fig. 7. Log(% Error) vs. ρ for 2/2/4 Queue

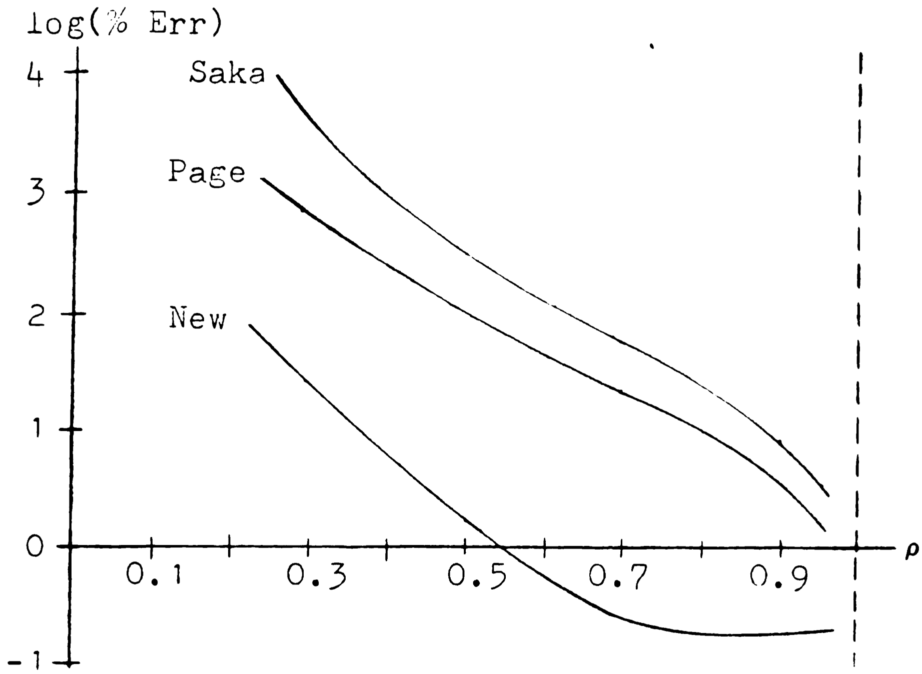

Fig. 8. Log(% Error) vs. ρ for 2/2/8 Queue

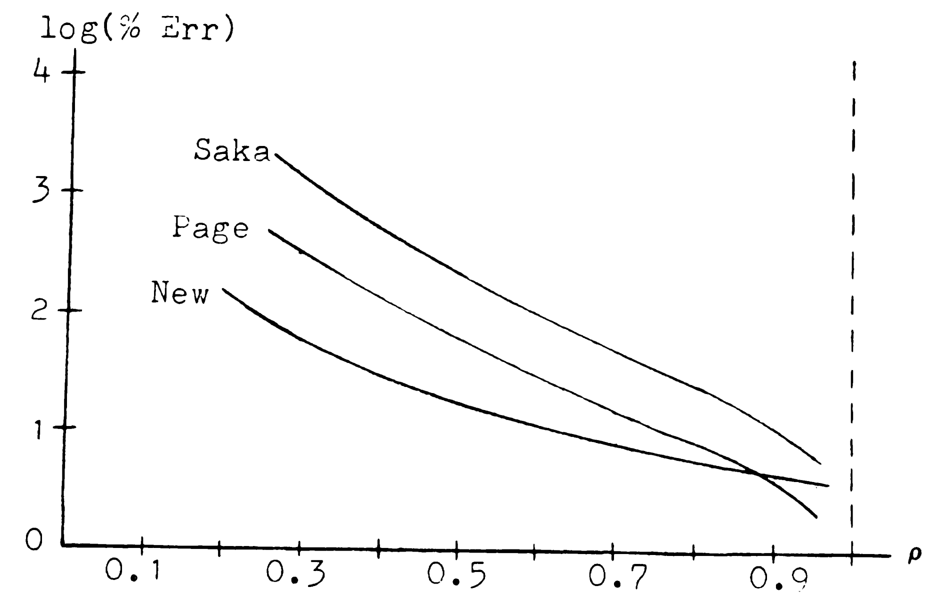

Fig. 9. Log(% Error) vs. ρ for 9/2/2 Queue

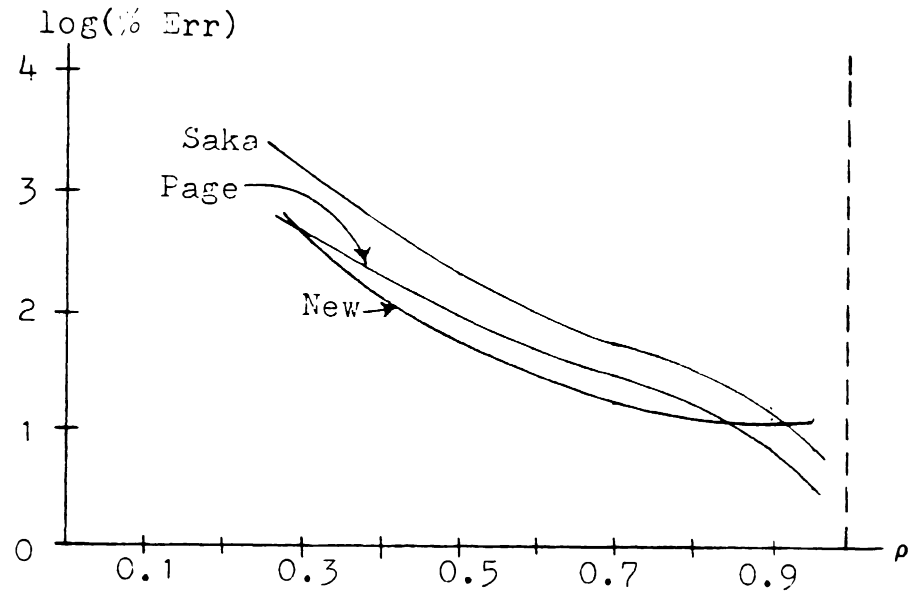

Fig. 10. Log(% Error) vs. ρ for 16/3/1 Queue

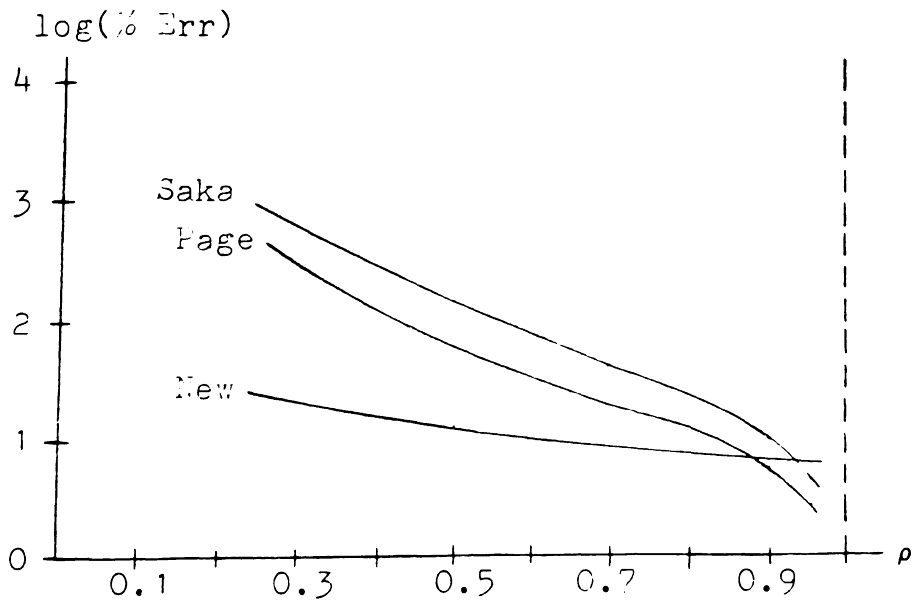

Fig. 11. Log(% Error) vs. ρ for 4/3/2 Queue

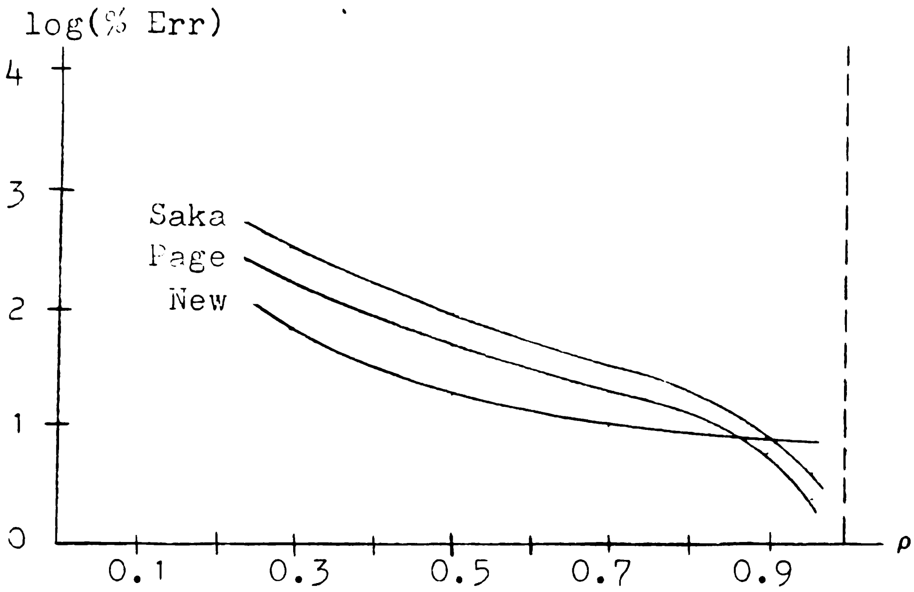

Fig. 12. Log(% Error) vs. ρ for 4/4/1 Queue

approaches one, like Page's and Sakasegawa's approximations. It is suspected that the new approximation would converge to exact values if we had used the calculated values of A and B in table 7 and the exact value of $W_{M/Ek/c}$ for the comparisons. We avoided this however, so that tables would not have to be used to apply the new approximation.

The biggest advantage of using Sakasegawa's Approximation is the relative ease of application. A hand-held calculator can be used for obtaining this rough estimate of waiting time. Page's Approximation requires the use of tables or a computer program, but is much more accurate. The new approximation requires the use of tables or a computer program, but gives the best results for moderate values of ρ.

VI. CONCLUSION

The error in the new approximation for waiting time comes from several sources. The basic limitation on accuracy is in the assumption that the A and B values of table 7 are independent of ρ. The errors caused by this assumption are comparable to those of table 6 for all systems tested.

The second source of error in the new approximation is using an approximation for $W_{M/Ek/c}$ rather than an exact solution. A better approximation for the M/Ek/c queue could improve the new approximation for the $E_m/E_k/c$ system. However the approximation used for the M/Ek/c queue is quite accurate so the error added by this is small.

The final and largest source of error in the new approximation comes from using eqs. 28 and 29 to calculate A and B rather than using the more accurate values in table 7. If eqs. 28 and 29 could be improved the new approximation could perform better than Page's Approximation for larger values of ρ. This is the primary area for future research on an approximation of this form.

As it stands the new approximation presented in this work produces more accurate results than any existing approximation for moderate values of ρ. To incorporate the new approximation into a computer program takes only about

100 lines of Fortran and a minimal amount of computer time (See Appendix II). Although it cannot be checked, it is suspected that the new approximation will perform well for values of m, k, and c outside the ranges of the tables of exact waiting times.

VII. REFERENCES

[1] D. Cox, and H. Miller, The Theory of Stochastic Processes, New York: Wiley, 1965.

[2] B. Halachmi, and W. R. Franta, "A diffusion approximation to the multi-server queue," Management Science 24, No. 2, p. 522, Jan. 1978.

[3] F. S. Hillier, and F. D. Lo, "Tables for multiple-server queueing systems involving Erlang distributions," Technical Report No. 31, Stanford University, Stanford California, Dec. 28, 1971.

[4] J. F. C. Kingman, "Inequalities in the theory of queues," Journal of the Royal Statistical Society, Series B, 32, p. 102, 1970.

[5] L. Kleinrock, Queueing Systems Volume 1: Theory, New York: Wiley, 1975.

[6] L. Kleinrock, Queueing Systems Volume 2: Computer Applications, New York: Wiley, 1976.

[7] H. Kobayashi, "Application of the diffusion approximation to queueing networks I: equilibrium queue distributions," J. ACM 21, 2, p. 316, April 1974.

[8] H. Kobayashi, Modeling and Analysis: An Introduction to System Performance Evaluation Methodology, Reading, Massachusetts: Addison-Wesley, 1981.

[9] J. Kollerstrom, "Heavy traffic theory for queues with several servers, I." Journal of Applied Probability, 11, p. 544, 1974.

[10] A. A. Nilsson, W. J. Stewart, and J. Seraj, "The analysis of multiprocessor systems by a diffusion approximation", Ninth International Teletraffic Conference, October 1979.

[11] E. Page, "Tables of waiting times for M/M/n, M/D/n, and D/M/n and their use to give approximate waiting times in more general queues," Journal of Operational Research Society, Vol. 33, p. 453, 1982.

[12] E. Page, Queueing Theory in OR, London: Butterworths, 1972.

[13] H. Sakasegawa, "An approximation formula $L = \alpha\rho^b/(1-\rho)$," Ann. Inst. Statist. Math 29, No. 1A, p. 67, 1977.

[14] L. Takacs, Introduction to the Theory of Queues, New York: Oxford University Press, 1962.

[15] Y. Takahashi, "An approximation formula for the mean waiting time of an M/G/c queue," Journal of the Operations Research Society of Japan, Vol. 20, No. 3, p. 150, Sept. 1977.

[16] W. Whitt, "The queueing network analyzer," The Bell System Technical Journal, Vol. 62, No. 9, p. 2779, Nov. 1983.

VIII. APPENDICES

Appendix I. Notation

Within this thesis the shorthand notation A/B/c is used to specify a given queueing system. A and B describe the interarrival time and service time distributions respectively. The number of parallel servers of the system is c.

The frequently used distributions are represented by the following symbols.

M Exponential distribution

D Deterministic or constant distribution

E_r r-stage Erlangian distribution

H_r r-stage Hyperexponential distribution

G General distribution

Thus the $E_2/M/2$ system has a 2-stage Erlangian interarrival time distribution, exponential service times, and 2 parallel servers. The $E_m/E_k/c$ system is sometimes referred to as the m/k/c system.

Appendix II. A Fortran Program for the New Approximation

```
c       Program to test New Approx. for
c       Em/Ek/c.  Offered load per
c       server varies 0.00 to 0.95.
c       Coefficients of var.LT one
        Print *,'m for arrival process='
        READ(5,*)CM
        CASQR=1.0/CM
        Print *,'k for service process='
        READ(5,*)CK
        CBSQR=1.0/CK
        Print *,'number of servers='
        READ(5,*)NSERV
        Print *,'utilization ='
        READ(5,*)RHO
        P=((1.0-CM)/CM)*((NSERV/10.0+1.1)/
      *(CM-1.0)+3.0-2.0/NSERV)
        AX=((1.0-CK)/CK)*exp(P)
        B1=0.0979*exp(-0.062*(CM-1)*(CM-1))
        B2=exp(-0.246*CK)
        B3=(0.409/NSERV)*exp(-0.0177*(CK-10)*(CK-10))
        BY=B1+B2+B3
        SIGMA=S(RHO,CM)
        X=WMG(CBSQR,NSERV,RHO)
        Y=WGM(CM,RHO,SIGMA,NSERV)
        WGG=abs(AX*X+BY*Y)
        WRITE(6,1000)WGG
 1000   FORMAT(5X,F10.5)
 100    END

        FUNCTION S(D,F)
c       D is rho, F is inverse of
c       casqr coef, function finds
c       unique sigma between 0 and
c       1 to solve for WEm/M/c
        ABOVE=1.0
        BELOW=0.0
        DO 5 N=1,7
        WIDTH=ABOVE-BELOW
        A=BELOW
        B=BELOW+0.9001*WIDTH
 110    A=A+0.1*WIDTH
        IF(A.GT.B) GOTO 11
        C=(D*F/(1.0-1.0*A+D*F))**F-A
        IF(C.LT.0.0) GOTO 11
```

```
      GOTO 110
 11   ABOVE=A
      BELOW=ABOVE-0.1*WIDTH
 5    CONTINUE
      S=ABOVE
      RETURN
      END

      FUNCTION WMG(B,N,P)
c     Finds approx. waiting time in M/Ek/c.
c     B is coef. of var. CBSQR,
c     N is number of servers, P is rho.
      A=N*P
      WMM=D(N,A)/(N-A)
      WMD=1.0+(1.0-P)*(N-1)*(SQRT(4.+5.*N)-2.0)
     */16.0/A
      WMD=WMD*WMM/2.0
      WMG=(1.0-B)*WMD+B*WMM
      RETURN
      END

      FUNCTION D(I,T)
c     Erlangs second formula, I is the number of
c     servers, T is the offered traffic
      D=1.0
      E=1.0
      IF(I.EQ.0) RETURN
      DO 1 J=1,I
 1    E=E*T/(J+T*E)
      D=I*E/(I-T*(1.0-E))
      RETURN
      END



      FUNCTION WGM(Z,RO,SIG,NSER)
c     Finds exact waiting time in Em/M/c.
c     z is inverse of coef, of var.
c     CASQR, RO is rho, SIG is sigma, (note mu=1)
c     NSER is number of servers.
      IF (NSER.EQ.1) GOTO 102
      SUM=0.0
      DO 3 J=1,NSER
      FIJ=(NSER*Z*RO/(NSER*Z*RO+J))**Z
      CJ=1.0
      DO 2 I=1,J
      BFI=(NSER*Z*RO/(NSER*Z*RO+I))**Z
      CJ=CJ*BFI/(1.0-BFI)
 2    CONTINUE
      NJ=NSER-J
```