AD-A042 722 STANFORD UNIV CALIF STANFORD ELECTRONICS LABS F/G 9/2

ON ACCURACY IMPROVEMENT AND APPLICABILITY CONDITIONS OF DIFFUSI--ETC(U)

JAN 77 P S YU DASG60-77-C-0073

UNCLASSIFIED SU-SEL-77-016 NL

1 OF 2

ADA042-722

DIGITAL SYSTEMS LABORATORY

STANFORD ELECTRONICS LABORATORIES
DEPARTMENT OF ELECTRICAL ENGINEERING
STANFORD UNIVERSITY · STANFORD, CA 94305

SEL-77-016

On Accuracy Improvement and Applicability Conditions of Diffusion Approximation with Applications to Modelling of Computer Systems

by

Philip S. Yu

January 1977

Technical Report No. 129

The work described herein was supported in part by the Ballistic Missile Defense Systems Command under contract no. DASG60-77-C-0073. Computer time was made available by the Stanford Linear Accelerator Center.

SECURITY CLASSIFICATION OF THIS PAGE (When Data

REPORT DOCUMENTATION P

READ INSTRUCTIONS BEFORE COMPLETING FORM

14

SU-DSL-TR-129

1. REPORT NUMBER: SEL-77-016, DSL Tech. Report # 129	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	9
4. TITLE (and Subtitle): 6 ON ACCURACY IMPROVEMENT AND APPLICABILITY CONDITIONS OF DIFFUSION APPROXIMATION WITH APPLICATIONS TO MODELLING OF COMPUTER SYSTEMS.	5. TYPE OF REPORT & PERIOD COVERED: Technical Report.
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): 10 Philip S. Yu	8. CONTRACT OR GRANT NUMBER(s): 15 DASG60-77-C-0073
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Stanford Electronics Laboratories, Stanford University, Stanford, CA 94305	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 6.33.04.A
11. CONTROLLING OFFICE NAME AND ADDRESS: Ballistic Missile Defense Advanced Technology Center, ATC-P , P.O. Box 1500, Huntsville, AL 35807	12. REPORT DATE: 11 Jan[redacted] 1977
	13. NUMBER OF PAGES: 98 12 109p.
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Reproduction in whole or in part is permitted for any purpose of the United States Government-

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Starting with single server queueing systems, we find a different way to timate the diffusion parameters. The boundary condition is handled using the ller's elementary return process. Extensive comparisons by asymptotic, ulation and numerical techniques have been conducted to establish the superiority the proposed method compared with conventional methods. The limitation of the fusion approximation is also investigated. When the coefficient of variation nterarrival time is larger than one, the mean queue length may vary over a wide even if the mean and variance of interarrival time are kept unchanged. The

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

332400

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Jmc

diffusion approximation is applicable under the condition that the high variation of interarrival time is due to a large number of short interarrival times. Case studies are conducted on 2-stage hyperexponential distributions. A similar anomaly is observed in two server closed queueing networks when the service time of any server has a large coefficient of variation. Again, a similar regularity condition on the service time distribution is required in order for the diffusion approximation to be applicable. For general queueing networks, the problems become more complicated. A simple way to estimate the coefficient of variation of interarrival time (when the network is decomposable) is proposed. Besides the anomalies cited before, networks under certain topologies, such as networks with feedback loops, especially self loops, can not be decomposed into separate single servers when the coefficient of variation of service time distributions become large, even if the large variations are due to a large number of short service times. Nevertheless, the decomposability of a network can be improved by replacing each server with a self loop by an equivalent server without a self loop. Finally, we consider the service center with a queue dependent service rate or arrival rate. Generalization to two server closed queueing networks where each server may have a self loop is also considered.

SEL-77-016

ON ACCURACY IMPROVEMENT AND APPLICABILITY CONDITIONS OF DIFFUSION APPROXIMATION WITH APPLICATIONS TO MODELLING OF COMPUTER SYSTEMS

by

Philip S. Yu

January 1977

Technical Report No. 129

Digital Systems Laboratory
Departments of Electrical Engineering and Computer Science
Stanford University
Stanford, CA 94305

The work described herein was supported in part by the Ballistic Missile Defense Systems Command under contract no. DASG60-77-C-0073. Computer time was made available by the Stanford Linear Accelerator Center.

SEL-77-016

Digital Systems Laboratory
Departments of Electrical Engineering and Computer Science
Stanford University
Stanford, CA 94305

Technical Report No. 129

January 1977

ON ACCURACY IMPROVEMENT AND APPLICABILITY CONDITIONS OF DIFFUSION APPROXIMATION WITH APPLICATIONS TO MODELLING OF COMPUTER SYSTEMS

by

Philip S. Yu

ABSTRACT

Starting with single server queueing systems, we find a different way to estimate the diffusion parameters. The boundary condition is handled using the Feller's elementary return process. Extensive comparisons by asymptotic, simulation and numerical techniques have been conducted to establish the superiority of the proposed method compared with conventional methods. The limitation of the diffusion approximation is also investigated. When the coefficient of variation of interarrival time is larger than one, the mean queue length may vary over a wide range even if the mean and variance of interarrival time are kept unchanged. The diffusion approximation is applicable under the condition that the high variation of interarrival time is due to a large number of short interarrival times. Case studies are conducted on 2-stage hyperexponential distributions. A similar anomaly is observed in two server closed queueing networks when the service time of any server has a large coefficient of variation. Again, a similar regularity condition on the service time distribution is required in order for the diffusion approximation to be applicable. For general queueing networks, the problems become more complicated. A simple way to estimate the coefficient of variation of interarrival time (when the network is decomposable) is proposed. Besides the anomalies cited before, networks under certain topologies, such as networks with feedback loops, especially self loops, can not be decomposed into separate single servers when the coefficient of variation of service time distributions become large, even if the large variations are due to a large number of short service times. Nevertheless, the decomposability of a network can be improved by replacing each server with a self loop by an equivalent server without a self loop. Finally, we consider the service center with a queue dependent service rate or arrival rate. Generalization to two server closed queueing networks where each server may have a self loop is also considered.

Acknowledgement: The author would like to thank Professor Michael J. Flynn and Dr. Hisashi Kobayashi for their helpful comments and suggestions.

The work described herein was supported in part by the Ballistic Missile Defense Systems Command under contract no. DASG60-77-C-0073. Computer time was made available by the Stanford Linear Accelerator Center.

CONTENTS

TABLES

TABLES (Cont)

TABLES (Cont)

ILLUSTRATIONS

1. INTRODUCTION

Recently, considerable effort has been made for obtaining approximate solutions to non-Markovian queueing models using the diffusion approximation when the traffic intensity of the queueing system is high. The advantage of diffusion approximation lies in the fact that explicit results can be obtained for relatively complex situations where the only possible alternatives are numerical methods or simulation experiments. This greatly extends our capability in modelling practical problems. In the past, over simplified models have often been used for the sake of mathematical tractability, and the predicted performance may sometimes be quite different from the actual measured performance.

In order to alleviate the difficulty involved with general service time distribution, the diffusion approximation replaces the discrete jump process such as the queue size process by a diffusion process which is a continuous path stochastic process. The probability distribution of the diffusion process which satisfies a partial differential equation is quite often more amenable to mathematical analysis than that of the jump process. However, the approximation by diffusion process requires the heavy traffic assumption, as we shall see in Section 2.

Based on central limit theorem, Kingman [8] has shown in his treatment of heavy traffic theory that the waiting time distribution is as an approximation exponentially distributed, where the parameter depends only on the mean and variance of the interarrival time and service time distribution, i.e., it is insensitive to the detailed form of the distribution, as the traffic intensity approaches 1. The diffusion approximation based on the same idea attempts to overcome the limitation of the exponential model by considering both the mean and variance of the service time and interarrival time distributions. Newell [11] gives an extensive treatment of queues with time dependent arrival rate through use of the diffusion approximation in his monograph. Gaver applies the diffusion approximation method to waiting time in a M/G/1 queue [4]. Gaver and Shedler [2,3] apply this technique to the analysis of a multiprogrammed computer system modelled as a two stage cyclic network. Kobayashi [10] considers the multi-dimensional diffusion approximation as a technique for treating general queueing networks. Reiser and Kobayashi [12] study the accuracy of diffusion approximation techniques and propose a way to treat each server in the queueing networks separately. Gelenbe [5,7] suggests a different way to handle the boundary condition of the diffusion process, namely using the Feller's elementary return process [1]. In [6], Gelenbe also investigates the idea of decomposing a queueing network into separate single servers. An application of the diffusion approximation to analyze the performance of an ALOHA-like system can be found in the paper by Kobayashi, Onozato, Huynch [17]. Kleinrock [9] also has a tutorial chapter on diffusion approximation.

Since the diffusion approximation to single server queueing systems serve as the foundation to the approximation of more complicated queueing networks, we will start with single server systems, then advance to two server closed queueing networks where each server may have a self loop, and finally examine the problem in general queueing networks.

In Section 2, we propose a new way to estimate the diffusion parameters. Using the Feller's elementary return process [1] as proposed by Gelenbe [5] to handle the boundary condition, the approximate mean queue length obtained by this method is more accurate than that by conventional methods in most cases, especially when the coefficient of variation of the service time is large. In Section 3, we analyze the asymptotic error in mean queue length by our method and two other widely used diffusion approximation techniques proposed by Kobayashi [10] and Gelenbe [5] for the M/G/1 and E_2/M/1 queueing systems, where analytic results on mean queue length are available in closed forms. The advantage of the asymptotic analysis is that the absolute or relative errors are expressed in terms of the traffic intensity or the coefficients of variation of the service time and interarrival time distributions. This provides better insight in understanding the accuracy of various approximation techniques. Kingman [18] has found a tight upper bound for the mean queue length. We also analyze this upper bound for reference. It is interesting to see that, in the M/G/1 system, the mean queue lengths obtained by the other two methods are larger than the Kingman's upper bound when the service time has a large coefficient of variation. In both E_2/M/1 and M/G/1 systems, our method yields more accurate approximations. In fact, in the M/G/1 system, the mean queue length obtained by our method is exact, and those obtained by the other two methods have an error term on the order of $C_s/2$ or $(C_s - 1)/2$, where C_s is the squared coefficient of variation of the service time. For readers who are not familiar with asymptotic analysis, this section can be skipped over. In Section 4, simulations have been conducted to test the relative performance of our method and the two conventional methods for more general queueing systems which include the $E_r/E_n/1$ and $E_r/H_2/1$ systems. Numerical techniques have also been employed to study the relative performance of various diffusion approximations for the $E_3/M/1$ and D/M/1 systems. Our method yields more accurate approximations, except in the $E_r/E_n/1$ system. In the $E_r/E_n/1$ system, the method in [4] proposed by Kobayashi has better performance than ours. These comprehensive and systematic comparisons not only establish the robustness of the proposed method, but also provide valuable information in selecting the best approximation technique for the specific problem at hand. In Section 5, we use the $H_2/M/1$ system to illustrate the fact that, when the coefficient of variation of the arrival process is larger than 1, the mean queue length may vary over a wide range even if the mean and variance of the interarrival time are kept unchanged. This is simply because the coefficient of variation of the distribution function may become large due to different reasons. We give a reasonable interpretation to this phenomenon. Since two-stage hyperexponential distribution function is widely used in computer system modelling, we try to identify the range of the parameters of the hyperexponential distribution where the diffusion approximation can be applied to obtain a fairly accurate estimation of the mean queue length under various traffic intensities. The data included in that section should be helpful in checking the applicability of diffusion approximation to the problem at hand. The $H_2/E_r/1$ and $H_2/H_2/1$ systems are also examined. In those cases where the parameters are not in the applicable range, the diffusion approximation may be used to estimate a lower bound of the system performance.

After examining the single server queueing system, in Section 6 we consider a more complicated queueing system, the closed two server system which is often used to model the computer system under fixed degree of multiprogramming, referred to as the CPU and DTU model [2,3]. The approximate utilization and mean queue length of the CPU are very close to the simulation or exact result when the coefficients of variation of the service time distributions are small [2,5]. When this condition does not hold, the diffusion approximation techniques can provide a close approximation to mean queue length and utilization only under restricted ranges of the parameters of the distribution functions and in other cases, it may still be used to estimate a lower bound of the performance of the system. Then, in Section 7, the decomposition problem of the network of queues is considered. From the data provided in Sections 3 and 4, the accuracy of diffusion approximation to a single server system is undoubtedly very good. The problem on decomposition of a queueing network into separate single server systems seems to be how to estimate the coefficient of variation of the interarrival time at each server, so we can take the interactions among interconnected servers into account. Two different methods have been proposed to estimate the coefficient of variation of the interarrival time at each server by Reiser and Kobayashi [12] and Gelenbe [6], respectively. Both methods lead to fairly accurate approximations. The second method which tries to incorporate the effect of idle periods on the coefficient of variation seems to be better but is more complicated. Here, we propose a method to estimate the coefficient of variation of the interarrival time which leads to similar results as the second method by taking the effect of idle periods into consideration but is much simpler in computation. All the examples given by the previous authors to demonstrate the accuracy on decomposing queueing networks into separate single servers under diffusion approximations are concentrated on the situation where the coefficients of variation of service time and external interarrival time distributions are not large, mainly less than or equal to two. Actually, the decomposition technique is not always feasible when the coefficients of variation of the service time or external interarrival time distributions become large. An example has been given to illustrate this anomaly which has been overlooked in the past. Hence, we must be careful on the decomposability of a queueing network. Although decomposability is an inherent property of the network topology, its effect magnifies as the coefficients of variation of service time distributions deviate largely from one. Nevertheless, decomposability of a network can be improved by replacing each server with a self loop by an equivalent server without a self loop. Finally, in Section 8, we consider the service center with a queue dependent service rate or arrival rate. Generalization to the closed two server queueing network is also considered.

2. THE DIFFUSION APPROXIMATION FOR THE G/G/1 QUEUE

Consider a single server queueing system. Let t_i be the arrival time of the i^{th} job to the queueing system and t_i' be the departure time of the i^{th} job, where $0 < t_1 < t_2 < \ldots,\ 0 < t_1' < t_2' < \ldots,$ and $t_i' > t_1$, i.e., the queueing discipline of the system is first come first served (FCFS). Let $A(t)$ and $D(t)$ represent the cumulative number of arrivals and departures, respectively, up to time t. Denote the number of jobs in the queue (including the job in service) at time t by $Q(t)$, then

$$Q(t) = A(t) - D(t)$$

Assume the interarrival time U_i's $(= t_i - t_{i-1})$ and service time V_i's are independent and identically distributed, respectively. Furthermore, we assume

$$E\{U_i\} = \frac{1}{\lambda}$$

$$Var\{U_i\} = \sigma_a^2$$

$$E\{V_i\} = \frac{1}{\mu}$$

$$Var\{V_i\} = \sigma_s^2$$

and

$$\rho = \frac{\lambda}{\mu}$$

where ρ is the traffic intensity of the queueing system.

The following central limit theorem for renewal processes [14] will be used in later discussion.

<u>Theorem</u>. If $T = \{T_n\}$ is a renewal process (i.e., $T_n - T_{n-1}$ are independent and identically distributed) for which

$$M = E\{T_1\} < \infty$$

and

$$\sigma^2 = E\left\{(T_1 - M)^2\right\} < \infty$$

Let

$$N(t) = \sum_{n=0}^{\infty} I_{[0,t]}(T_n)$$

where

$$I_A(X) = \begin{cases} 1 & \text{if} \quad X \in A \\ 0 & \text{if} \quad X \notin A \end{cases}$$

i.e., $N(t)$ is the number of T_n such that $T_n \leq t$. Then

$$\lim_{t \to \infty} P\left\{\frac{N(t) - t/M}{\sqrt{t\sigma^2/M^3}} < X\right\} = \Phi(X)$$

where

$$\Phi(X) = \frac{1}{2\pi} \int_{-\infty}^{X} \exp\left(-\frac{1}{2} u^2\right) du$$

is the normal integral.

Now let us introduce the definition and property of a diffusion process.

Definitions. A diffusion process $\{X(t), t \geq 0\}$ is a strong Markov process such that

(1) $\beta(X,t) = \lim_{\Delta t \to 0} \dfrac{E\{X(t + \Delta t) - X(t) \mid X(t) = X\}}{\Delta t}$ exists

(2) $\alpha(X,t) = \lim_{\Delta t \to 0} \dfrac{E\{(X(t + \Delta t) - X(t))^2 \mid X(t) = X\}}{\Delta t}$ exists

(3) the sample path is continuous

where the diffusion parameters $\beta(X,t)$ and $\alpha(X,t)$ are called the infinitesimal mean and variance coefficients, respectively.

Let $P(X_0,X;t)$ be the probability density function of the diffusion process $X(t)$, i.e.,

$$P(X_0,X;t)\,dX = P\left\{X \le X(t) \le X + dX \,|\, X(0) = X_0\right\}$$

then $P(X_0,X;t)$ will satisfy the following differential equation:

$$\frac{\partial}{\partial t} P(X_0,X;t) = \frac{1}{2}\frac{\partial}{\partial X}\ \alpha(X,t)\ \frac{\partial}{\partial X} P(X_0,X;t) - \frac{\partial}{\partial X}\beta(X,t)\ P(X_0,X;t)$$

which is called the Kolmogorov diffusion equation or Fokker-Plank equation [14].

Clearly, as t becomes large, the renewal counting process $N(t)$ is approaching a diffusion process with

$$\beta(X,t) = \frac{1}{M}$$

and

$$\alpha(X,t) = \frac{\sigma^2}{M^3}$$

which is usually referred to as the Wiener process with drift.

By assumption, the arrival process is a renewal process, hence $A(t)$ will converge to a Wiener process with infinitesimal mean λ and infinitesimal variance $\sigma_a^2\lambda^3$, as t becomes large. The problem is that the interdeparture time is not independent and identically distributed, since the interdeparture time can either be a service time or the sum of a service time and an idle period of the server. Hence, the departure process is not a renewal process. But, under heavy traffic conditions, i.e., as $\rho \to 1$, it is close to a renewal process. During the busy period, $D(t)$ will come close to a Wiener process with infinitesimal mean μ and variance coefficient $\sigma_s^2\mu^3$ as t increases, provided that the busy period is not interrupted. Still another problem is that $A(t)$ and $D(t)$ is not independent since $Q(t) = A(t) - D(t) \ge 0$. But, when $Q(t)$ is larger than zero, we have a departure process independent of the arrival process. In this case, $Q(t)$ behaves like a Wiener process which is a diffusion process with no boundary restriction at zero. That is to say, we should approximate $Q(t)$ by a diffusion process with appropriate boundary condition at zero to reflect the fact that $Q(t)$ can never become negative and there is an idle period after $Q(t)$ drops to zero.

To be more precise, $Q(t)$ will converge to a diffusion process with parameters

$$\beta(X,t) = \lambda - \mu$$

$$\alpha(X,t) = \sigma_a^2 \lambda^3 + \sigma_s^2 \mu^3 \rho$$

The parameters are obtained from those of $A(t)$ and $D(t)$ based on the fact that $Q(t) = A(t) - D(t)$. An extra factor ρ appeared in the second term of $\alpha(X,t)$ is used to reflect the fact that $D(t)$ has a coefficient of variation $\sigma_s^2 \mu^3$ only ρ of the time.

Since both $\beta(X,t)$ and $\alpha(X,t)$ are constant, we will abbreviate them as β and α, respectively. The probability density function $P(X_0,X;t)$ of $Q(t)$ will satisfy the equation

$$\frac{\partial}{\partial t} P(X_0,X,t) = \frac{\alpha}{2} \frac{\partial^2}{\partial X^2} P(X_0,X,t) - \beta \frac{\partial}{\partial X} P(X_0,X,t)$$

Let $P(X)$ be the stationary density function of $Q(t)$, i.e.,

$$P(X)\, d(X) = P\{X \le Q(t) \le X + dX\} \qquad \text{as} \quad t \to \infty$$

For the stationary case, the time derivative in the Fokker-Plank equation is set to zero. So

$$\frac{\alpha}{2} \frac{\partial^2}{\partial X^2} P(X) - \beta \frac{\partial}{\partial X} P(X) = 0 \tag{2.1}$$

Two different approaches have been suggested to handle the boundary condition. The first approach is to treat the boundary $X = 0$ as a reflecting boundary, i.e., whenever the queue becomes empty, it is reflected to positive immediately. Though the queue size will never become negative, still no probability mass can collect at $X = 0$. Gaver and Shedler

[2,3] and Kobayashi [10], who generalized this approach to queueing network, have managed to choose the appropriate integration constants in the solution to (2.1) under reflecting boundary, so that the model correctly predicts the stationary probability of empty queue. The second approach proposed by Gelenbe [5,6,7] uses Feller's elementary return process [1] instead of the diffusion process with reflecting boundary to approximate the queueing system. This is a diffusion process with boundary to which the process adheres for epochs whenever the process attains a boundary; at the end of the epoch the process is reinitialized according to a fixed probability density function. Gelenbe [5] first solves the equation when the holding time on the boundary has exponential distribution and later on [7] generalizes it to any probability density function whose Laplace-Stieltjes transform is a rational function to account for the fact that the holding time in the boundary in general is not exponentially distributed. Fortunately, the solution under the general distribution depends only on the first moment of the holding time distribution. Thus, the stationary solution is identical to the corresponding solution when the holding time is exponentially distributed with the same mean.

In this paper, we will adopt Gelenbe's approach to handle the boundary condition and assume exponential holding time on the boundary, since this assumption will simplify the problem and lead to the same solution as that under general holding time distribution [7]. The advantage of this approach is that it can be extended very easily to handle two server closed queueing networks or finite capacity queues. Both of them have important applications in computer modelling. The only problem we are facing is that the mean holding time, h, at the boundary $X = 0$ is not known. The holding time at the boundary $X = 0$ is, in fact, the idle period of the queueing system. From queueing theory [23], we know that

$$\begin{aligned} h &= E\{\text{idle period}\} \\ &= \lambda^{-1}(1 - \rho)\, E\{n\} \end{aligned}$$

where $E\{n\}$ is the expected number of jobs being served in each busy period, and furthermore

$$E\{n\} = \exp \sum_{k=1}^{\infty} \frac{1}{k} P\{S_k > 0\}$$

where

$$S_0 = 0$$

$$S_n = \sum_{i=1}^{n} (V_{i-1} - U_i)$$

Recall V_i is the service time of the i^{th} customer and U_i is the interarrival time between the i^{th} and $i\text{-}1^{th}$ customer. The expression for $E\{n\}$ can be simplified to $1/(1-\rho)$ when the interarrival time, U_i, has exponential distribution, i.e., the system is M/G/1. In general, it can not be simplified. We will use the conditions that the integration of probability density function over the range $X \geq 0$ should equal to one to obtain an estimation of the holding time, h. To account for the fact that, after an arrival to the empty queue occurs the number of customers in the queue jumps instantaneously to one, we need to add an extra term, $-(1-\rho)/h\ \delta(X-1)$, to the right hand side of (2.1) and an extra boundary equation (2.3), as explained below.

Now we have the following equations

$$\frac{\alpha}{2} \frac{\partial^2}{\partial X^2} P(X) - \beta \frac{\partial}{\partial X} P(X) = - \frac{(1-\rho)}{h} \delta(X-1) \tag{2.2}$$

and

$$\lim_{X \to 0} \left[\frac{\alpha}{2} \frac{\partial}{\partial X} P(X) - \beta P(X)\right] = \frac{(1-\rho)}{h} \tag{2.3}$$

where $\delta(X-1)$ is a Dirac density function concentrated at $X = 1$ and represents the probability density function of the point from which the diffusion process starts once again immediately after a jump. Notice $(1-\rho)/h$ is the product of the probability at the boundary and the rate of jumping back from the boundary, i.e., $(1-\rho)/h$ represents the mean rate of jumping back to $(0,\infty)$. $(\alpha/2)(\partial/\partial X)\ P(X) - \beta P(X)$ has the physical interpretation as the rate of flow of the probability mass from the region $(0,\infty)$ to the boundary 0. This explains the boundary equation.

Similar arguments can be given to the correction term in (2.2). For a more detailed argument, see [5,7].

Let us denote

$$r = \frac{2\beta}{\alpha} = \frac{-2(\mu - \lambda)}{\sigma_a^2\lambda^3 + \sigma_s^2\mu^3\rho} = \frac{-2(1 - \rho)}{\rho(C_a + C_s)}$$

where

$$C_a = \sigma_a^2\lambda^2$$

$$C_s = \sigma_s^2\mu^2$$

C_a and C_s are called the squared coefficient of variations of the interarrival time and service time, respectively.

Solving the differential equation (2) with boundary condition (3) and $\lim_{X\to 0} P(X) = 0$, we get

$$P(X) = \begin{cases} \frac{(1 - \rho)}{h\beta}(e^{rX} - 1) & 0 \leq X \leq 1 \\ \frac{(1 - \rho)}{h\beta}[1 - e^{-r}]\, e^{rX} & X \geq 1 \end{cases}$$

To compute h, we use the fact that

$$(1 - \rho) + \int_0^\infty P(X)\, dX = 1$$

and yield, when $r < 0$ (i.e., $\rho < 1$),

$$h = \frac{1}{1 - \rho}$$

As we can see, the estimation of the length of idle period by diffusion approximation is only exact for the M/G/1 system.

Finally, we face the problem of discretization of the probability density function in the neighborhood of integer valued points $X = i$ in order to approximate π_i, the stationary probability of finding i customers in the queueing system. Usually there are three choices for π_i:

(1) $\int_{i-1}^{i} P(X)\,dX$, (2) $\int_{i}^{i+1} P(X)\,dX$, (3) $\int_{i-1/2}^{i+1/2} P(X)\,dX$. Here, we choose the first alternate, since it leads to more accurate approximation than the others under our method.

Let

$$\pi_0 = 1 - \rho$$

$$\pi_i = \int_{i-1}^{i} P(X)\,dX \qquad \text{for } i > 0$$

We get

$$\begin{cases} \pi_0 = 1 - \rho \\ \pi_1 = \dfrac{\rho^2(C_s + C_a)}{2(1-\rho)} (e^r - 1 - r) \\ \pi_i = \dfrac{\rho^2(C_s + C_a)}{2(1-\rho)} e^{ir}(1 - e^{-r})^2 \qquad i \geq 2 \end{cases} \tag{2.4}$$

The mean queue length under stationary distributions is

$$E\{Q\} = \sum_{i=1}^{\infty} i\pi_i = \rho\left[1 + \frac{\rho(C_s + C_a)}{2(1-\rho)}\right] \tag{2.5}$$

3. ACCURACY ANALYSIS OF DIFFUSION APPROXIMATION BY ASYMPTOTIC TECHNIQUE

In this section, we analyze the asymptotic errors of the mean queue lengths obtained by various diffusion approximation techniques for the M/G/1 system and the E_2/M/1 system where analytic solution of the mean queue length is available as $\rho \to 1$. Clearly, it is an important requirement for an approximation technique which are designed to handle general distributions for service time or interarrival time to give accurate approximation on those cases for which solutions are known. These error analyses certainly have important implications on the accuracy of the approximation for queueing systems whose service time and interarrival time distributions do not deviate too far from those of the M/G/1 or E_2/M/1 system. The advantage of asymptotic analysis is that we can obtain a closed form expression for the error term, and the order of the error can be clearly expressed in terms of the power of $(1-\rho)$ which gives us a clear picture of the dependency on heavy traffic assumption. Various diffusion approximation techniques have been proposed to handle the single server systems. The two most noteworthy methods are proposed by Reiser and Kobayashi [10] and Gelenbe [6], respectively. Since there is not any comprehensive study of the relative accuracy of the two methods, we will analyze not only the proposed method but also the two methods mentioned above. Our method improves the accuracy in both cases. The mean queue length obtained under Kingman heavy traffic approximation [8] has been proved to provide an upper bound on mean queue length [18]. The result holds for $0 \leq \rho < 1$ and improves to be a tight upper bound as $\rho \to 1$. The upper bound will also be analyzed for comparison. And we shall see this upper bound is quite tight in both cases. From then on, we will denote

(1) Method P: the proposed method

(2) Method A: the diffusion approximation technique proposed by Kobayashi [10]

In this approximation method, we have

$$\alpha = \sigma_a^2\lambda^3 + \sigma_s^2\mu^3$$

$$\beta = \mu - \lambda$$

The boundary at $X = 0$ is treated as a reflection boundary. By setting the probability mass at the origin to be $1 - \rho$ and solving the Fokker-Plank equation (2.1), the following stationary queue length distribution and mean queue length is obtained.

(a) Stationary queue length distribution

$$\pi_0 = 1 - \rho$$

$$\pi_i = \rho(1 - \hat{\rho})\, \hat{\rho}^{\,i-1} \qquad \text{for } i \geq 1 \tag{3.1}$$

where

$$\hat{\rho} = e^{-\frac{2(1-\rho)}{C_s + \rho C_a}}$$

(b) Stationary mean queue length

$$E\{Q_A\} = \frac{\rho}{1 - \hat{\rho}} \tag{3.2}$$

(3) Method B: the diffusion approximation technique proposed by Gelenbe [5]

As noted earlier, method P follows the argument in method B to handle the boundary condition, but the diffusion parameters in the two methods are different. The diffusion parameters in method B are

$$\alpha = \sigma_a^2 \lambda^3 + \sigma_s^2 \mu^3$$

$$\beta = \mu - \lambda$$

Hence, the stationary queue length distribution is

$$\pi_0 = 1 - \rho$$

$$\pi_1 = \frac{\rho(\rho C_a + C_s)}{2(1 - \rho)} (e^r - 1 - r) \tag{3.3}$$

$$\pi_i = \frac{\rho(\rho C_a + C_s)}{2(1-\rho)} e^{ir}(1 - e^{-r})^2 \qquad i \geq 2 \tag{3.3 Cont.}$$

where

$$r = \frac{2\beta}{\alpha} = \frac{2(\mu - \lambda)}{\sigma_a^2 \lambda^3 + \sigma_s^2 \mu^3}$$

and the mean queue length is

$$E\{Q_B\} = \rho\left(1 + \frac{\rho C_a + C_s}{2(1-\rho)}\right) \tag{3.4}$$

Notice the Kingman's upper bound on mean queue length [18] is

$$E\{Q_K\} = \frac{C_a + C_s \rho^2}{2(1-\rho)} + \rho \tag{3.5}$$

As we shall see, in the M/G/1 system the mean queue length obtained by method P is exact and those obtained by methods A and B have an absolute error around $\rho(C_s - 1)/2$ and $\rho C_s/2$, respectively. Hence, the performance of methods A and B will degrade as the coefficient of variation of the service time increases, e.g., when C_s equals 64 and ρ equals 0.8, the relative errors of both methods are around 25%. In fact, under heavy traffic condition the mean queue lengths obtained by methods A and B are larger than the Kigman's upper bound on mean queue length when C_s is larger than 3. In the E_2/M/1 system, again, the mean queue length obtained by method P is more accurate. The result from method A is very close to that from method P. Examining the asymptotic expressions for the mean queue lengths of both methods, we find their difference is proportional to $(1 - \rho)$. The mean queue length obtained by method B is less accurate and is quite close to the Kingman's upper bound on mean queue length.

We will first prove the following lemma, which is the foundation of the analysis on method A.

Lemma 1. The asymptotic mean queue length obtained under method A will satisfy

$$E\{Q_A\} = \frac{\rho}{1-\hat{\rho}}$$

$$= \frac{\rho(C_s + \rho C_a)}{2}(1-\rho)^{-1} + \frac{\rho}{2} + \frac{\rho}{6(C_s + \rho C_a)}(1-\rho) + 0\left((1-\rho)^2\right) \tag{3.6}$$

where

$$\hat{\rho} = e^{-\frac{2(1-\rho)}{C_s + \rho C_a}} \tag{3.7}$$

Proof.

Using the Taylor series expansion for e^X, i.e.,

$$e^{-X} = 1 - X + \frac{X^2}{2!} - \frac{X^3}{3!} + 0(X^4) \qquad \text{as} \quad X \to 0$$

we obtain, as $\rho \to 1$, from (3.7)

$$1 - \hat{\rho} = 1 - e^{-\frac{2(1-\rho)}{C_s + \rho C_a}}$$

$$= 1 - \left(1 - \frac{2(1-\rho)}{C_s + \rho C_a} + \frac{2(1-\rho)^2}{(C_s + \rho C_a)^2} + \frac{4(1-\rho)^3}{3(C_s + \rho C_a)^3} + 0\left((1-\rho)^4\right)\right)$$

$$= \frac{2(1-\rho)}{C_s + \rho C_a}\left(1 - \frac{1-\rho}{C_s + \rho C_a} + \frac{2(1-\rho)^2}{3(C_s + \rho C_a)^2} + 0\left((1-\rho)^3\right)\right) \tag{3.8}$$

Using the Taylor series expansion of

$$\frac{1}{1-X} = 1 + X + X^2 + 0(X^3) \qquad \text{as} \quad X \to 0$$

we obtain, as $\rho \to 1$, from (3.8)

$$\frac{\rho}{1-\hat{\rho}} = \frac{\rho(C_s + \rho C_a)}{2(1-\rho)} \frac{1}{\left(1 - \left(\frac{1-\rho}{C_s + \rho C_a} - \frac{2(1-\rho)^2}{3(C_s+\rho C_a)^2} + 0\left(\left(1-\rho\right)^3\right)\right)\right)}$$

$$= \frac{\rho(C_s + \rho C_a)}{2(1-\rho)} \left(1 + \frac{(1-\rho)}{C_s + \rho C_a} - \frac{2}{3}\frac{(1-\rho)^2}{(C_s+\rho C_a)^2} + \frac{(1-\rho)^2}{(C_s+\rho C_a)^2} + 0\left((1-\rho)^3\right)\right)$$

$$= \frac{\rho(C_s + C_a)}{2}(1-\rho)^{-1} + \frac{\rho}{2} + \frac{\rho}{6(C_s+\rho C_a)}(1-\rho) + 0\left((1-\rho)^2\right)$$

3.1 Mean Queue Length for M/G/1 System

The mean queue length of the M/G/1 system is given by the well known Pollaczek-Khintchine formula

$$E\{Q\} = \rho\left(1 + \frac{\rho}{2}\frac{(1+C_s)}{1-\rho}\right)$$

The mean queue length obtained by the proposed method is given by (2.5). Setting C_a equal to one, since the arrival process is Poisson, we find that the mean queue length given in (2.5) is exactly the Pollaczek-Khintchine formula. That is to say, we predict the mean queue length exactly for the M/G/1 system. Let us examine the asymptotic performance of method A and method B in this case. From Lemma 1, we get the mean queue length under method A, by setting $C_a = 1$,

$$E\{Q_A\} = \frac{\rho(C_s + \rho)}{2}(1-\rho)^{-1} + \frac{\rho}{2} + \frac{\rho}{6(C_s+\rho)}(1-\rho) + 0\left((1-\rho)^2\right)$$

and the absolute error is

$$E\{Q_A\} - E\{Q\} = \frac{\rho}{2}(C_s - 1) + \frac{\rho}{6(C_s + \rho)}(1 - \rho) + 0\left((1 - \rho)^2\right)$$

From equation (3.4), we get the mean queue length under method B, by setting $C_a = 1$,

$$E\{Q_B\} = \rho\left[1 + \frac{(\rho + C_s)}{2(1 - \rho)}\right]$$

and the absolute error is

$$E\{Q_B\} - E\{Q\} = \frac{\rho C_s}{2}$$

Since

$$E\{Q_B\} - E\{Q_A\} = \frac{\rho}{2} - \frac{\rho}{6(C_s + \rho)}(1 - \rho) + 0\left((1 - \rho)^2\right)$$

method A and method B have similar performance when C_s is large.

From equation (3.5), we get the Kingman's upper bound on mean queue length, by setting $C_a = 1$,

$$E\{Q_K\} = \frac{1}{2}\,\frac{1 + C_s\rho^2}{(1 - \rho)} + \rho$$

and the absolute error is

$$E\{Q_k\} - E\{Q\} = \frac{1}{2}(1 + \rho)$$

As we can see, the Kingman's upper bound is quite acceptible under the M/G/1 case and differs from the mean queue length by $1/2(1 + \rho)$. Although the absolute errors under both method A and method B become large as the coefficient of variation of the service time increases, the absolute error in the Kingman's upper bound is fixed. For $C_s > 3$, the mean queue length obtained by method A and method B is, in fact, larger than the Kingman's upper bound when the traffic is high. Let us examine the relative errors in both methods A and B as C_s becomes large.

The relative error for method A is

$$\frac{E\{Q_A\} - E\{Q\}}{E\{Q\}} = \frac{\frac{\rho}{2}(C_s - 1) + \frac{\rho}{6(C_s + \rho)}(1-\rho) + 0\left((1-\rho)^2\right)}{\frac{\rho^2}{2}\frac{(1+C_s)}{1-\rho}\left(1 + \frac{2(1-\rho)}{\rho(1+C_s)}\right)}$$

$$= \frac{2(1-\rho)}{\rho^2(1+C_s)}\left(\frac{\rho}{2}(C_s - 1) + \frac{\rho}{6(C_s+\rho)}(1-\rho) + 0\left((1-\rho)^2\right)\right)$$

$$\cdot \left(1 - \frac{2(1-\rho)}{\rho(1+C_s)} + 0\left((1-\rho)^2\right)\right)$$

$$= \frac{2(1-\rho)}{\rho^2(1+C_s)}\left(\frac{\rho}{2}(C_s - 1) + \left(\frac{\rho}{6(C_s+\rho)} - \frac{C_s - 1}{C_s + 1}\right)(1-\rho)\right) + 0\left((1-\rho)^3\right)$$

$$= \frac{C_s - 1}{\rho(C_s + 1)}(1-\rho) + \frac{2\left(7\rho - (5\rho - 6)C_s - 6C_s^2\right)}{\rho^2(1+C_s)^2(\rho + C_s)}(1-\rho)^2 + 0\left((1-\rho)^3\right)$$

and similarly the relative error for method B is

$$\frac{E\{Q_B\} - E\{Q\}}{E\{Q\}} = \frac{C_s}{\rho(C_s + 1)}(1-\rho) - \frac{2C_s}{\rho^2(1+C_s)^2}(1-\rho)^2 + 0\left((1-\rho)^3\right)$$

As C_s becomes large, the relative errors for both methods A and B approach $(1-\rho)/\rho$. Hence, for ρ equals 0.8, the relative errors are 25% as mentioned earlier (see Table 3.1). In Tables 3.1a and 3.1b, some numerical comparisons of methods P, A, and B are presented for $C_s = 0$, 1/5, 1/4, 1/3, 1/2, 1, 2, 4, 8, 16, 32, 64, and 128, when $\rho = 0.9$ and 0.8, respectively. This can be used as a check for the correctness of the asymptotic analysis.

Table 3.1a

MEAN QUEUE LENGTH FOR M/G/1 SYSTEM WHEN $\rho = 0.9$

C_s	Method P (Exact Result)	Method B	Method A
128	523.35	580.95	580.50
64	264.15	292.95	292.50
32	134.55	148.95	148.50
16	69.75	76.95	76.50
8	37.35	40.95	40.50
4	21.15	22.95	22.50
2	13.05	13.95	13.51
1	9.00	9.45	9.01
1/2	6.97	7.20	6.76
1/3	6.30	6.45	6.01
1/4	5.96	6.07	5.64
1/5	5.76	5.85	5.41
0	4.95	4.95	4.52

Table 3.1b

MEAN QUEUE LENGTH FOR M/G/1 SYSTEM WHEN $\rho = 0.8$

C_s	Method P (Exact Result)	Method B	Method A
128	207.20	258.40	258.00
64	104.80	130.40	130.00
32	53.60	66.40	66.00
16	28.00	34.40	34.00
8	15.20	18.40	18.00
4	8.80	10.40	10.00
2	5.60	6.40	6.00
1	4.00	4.40	4.01
1/2	3.20	3.40	3.02
1/3	2.93	3.07	2.69
1/4	2.80	2.90	2.53
1/5	2.72	2.80	2.43
0	2.40	2.40	2.03

3.2 Mean Queue Length for $E_2/M/1$ System

In the G/M/1 queueing system, the stationary distribution of number of customers in the system is given by [9]:

$$\begin{cases} \pi_0 = (1 - \sigma) \\ \pi_k = \rho(1 - \sigma)\,\sigma^{k-1} \qquad k \geq 1 \end{cases} \tag{3.9}$$

where σ is the unique root of

$$\sigma = A^*(\mu - \mu\sigma) \tag{3.10}$$

and $A^*(s)$ is the Laplace Stieltjes transform of the interarrival time distribution. Furthermore, the mean queue length is given by

$$E\{Q\} = \frac{\rho}{1 - \sigma} \tag{3.11}$$

When the interarrival time distribution is the 2-stage Erlang distribution, we can get a closed form solution for σ, i.e.,

$$\sigma = \frac{4\rho + 1 - \sqrt{8\rho + 1}}{2} \tag{3.12}$$

Before we proceed to compare the asymptotic behavior of various diffusion approximation techniques, we will first prove the following lemma.

Lemma. The steady state mean queue length of $E_2/M/1$ system will satisfy

$$E\{Q\} = \frac{3}{4}\,\rho(1 - \rho)^{-1} + \frac{\rho}{12} + \frac{5}{108}(1 - \rho) + 0\left((1 - \rho)^2\right) \qquad \text{as } \rho \to 1$$

Proof.

From (3.12),

$$1 - \sigma = \frac{1}{2}(1 - 4\rho + \sqrt{8\rho + 1})$$

Setting $Z = 1 - \rho$, after simplification, we get

$$1 - \sigma = \frac{1}{2}\left(4Z - 3 + 3\sqrt{1 - \frac{8}{9}Z}\right)$$

Using the Taylor's series expansion of $\sqrt{1 - X}$, we get

$$1 - \sigma = \frac{1}{2}\left(4Z - 3 + 3\left(1 - \frac{4}{9}Z - \frac{8}{81}Z^2 - \frac{32}{729}Z^3 + 0(Z^4)\right)\right)$$

$$= \frac{4}{3}Z\left(1 - \frac{1}{9}Z - \frac{4}{81}Z^2 + 0(Z^3)\right) \qquad (3.13)$$

Combining (3.12) and (3.13) together, we get

$$E\{Q\} = \frac{\rho}{\frac{4}{3}Z\left(1 - \frac{1}{9}Z - \frac{4}{81}Z^2 + 0(Z^3)\right)}$$

Using the Taylor series expansion of $(1 - X)^{-1}$, we get

$$E\{Q\} = \frac{3\rho}{4Z}\left(1 + \left(\frac{1}{9}Z + \frac{4}{81}Z^2 + 0(Z^3)\right) + \left(\frac{1}{9}Z + \frac{4}{81}Z^2 + 0(Z^3)\right)^2 + 0(Z^3)\right)$$

$$= \frac{3\rho}{4Z}\left(1 + \frac{1}{9}Z + \frac{5}{81}Z^2 + 0(Z^3)\right)$$

$$= \frac{3\rho}{4Z} + \frac{\rho}{12} + \frac{5}{108}\rho Z + 0(Z^2)$$

Finally, substituting $1 - \rho$ for Z, we get

$$E\{Q\} = \frac{3}{4}\rho(1 - \rho)^{-1} + \frac{\rho}{12} + \frac{5}{108}\rho(1 - \rho) + 0\left((1 - \rho)^2\right)$$

From equation (2.5), we get the mean queue length under the proposed method P, by setting $C_a = 1/2$, $C_s = 1$,

$$E\{Q_P\} = \rho + \frac{3}{4}\rho^2(1 - \rho)^{-1}$$

and the absolute error is

$$E\{Q_P\} - E\{Q\} = \frac{1}{6}\rho - \frac{5}{108}\rho(1-\rho) + 0\left((1-\rho)^2\right)$$

Furthermore, the relative error is

$$\frac{E\{Q_P\} - E\{Q\}}{E\{Q\}} = \frac{\frac{1}{6}\rho - \frac{5}{108}\rho(1-\rho) + 0\left((1-\rho)^2\right)}{\frac{3}{4}(1-\rho)^{-1} + \frac{\rho}{12} + 0\left((1-\rho)\right)}$$

$$= \frac{\frac{1}{6}\rho - \frac{5}{108}\rho(1-\rho) + 0\left((1-\rho)^2\right)}{\frac{3}{4}\rho(1-\rho)^{-1}}\left(1 - \frac{1}{9}(1-\rho) + 0\left((1-\rho)^2\right)\right)$$

$$= \frac{2}{9}(1-\rho) - \frac{7}{81}(1-\rho)^2 + 0\left((1-\rho)^3\right)$$

From Lemma 1, we get the mean queue length under method A, by setting $C_a = 1/2$, $C_s = 1$,

$$E\{Q_A\} = \frac{\rho\left(1 + \frac{1}{2}\rho\right)}{2}(1-\rho)^{-1} + \frac{\rho}{2} + \frac{\rho(1-\rho)}{6\left(1 + \frac{1}{2}\rho\right)} + 0\left((1-\rho)^2\right)$$

and the absolute error is

$$E\{Q_A\} - E\{Q\} = \frac{1}{6}\rho + \left(\frac{1}{6\left(1 + \frac{1}{2}\rho\right)} - \frac{5}{108}\right)\rho(1-\rho) + 0\left((1-\rho)^2\right)$$

Furthermore, the relative error is

$$\frac{E\{Q_A\} - E\{Q\}}{E\{Q\}} = \frac{2}{9}(1-\rho) + \left(\frac{2}{9\left(1 + \frac{\rho}{2}\right)} - \frac{7}{81}\right)(1-\rho)^2 + 0\left((1-\rho)^3\right)$$

From equation (3.4), we get the mean queue length under method B, by setting $C_a = 1/2$, $C_s = 1$,

$$E\{Q_B\} = \left(\frac{1}{4}\rho^2 + \frac{1}{2}\rho\right)(1-\rho)^{-1} + \rho$$

and the absolute error is

$$E\{Q_B\} - E\{Q\} = \frac{2}{3}\rho - \frac{5}{108}\rho(1-\rho)$$

Furthermore, the relative error is

$$\frac{E\{Q_B\} - E\{Q\}}{E\{Q\}} = \frac{8}{9}(1-\rho) - \frac{14}{81}(1-\rho)^2 + 0\left((1-\rho)^3\right)$$

From equation (3.5), we get the Kingman upper bound on mean queue length, by setting $C_a = 1/2$, $C_s = 1$,

$$E\{Q_k\} = \left(\frac{1}{4} + \frac{1}{2}\rho^2\right)(1-\rho)^{-1} + \rho$$

and the absolute error is

$$E\{Q_k\} - E\{Q\} = \left(\frac{1}{4} + \frac{5}{12}\rho\right) - \frac{5}{108}\rho(1-\rho) + 0\left((1-\rho)^2\right)$$

From the above analysis, it is clear that the mean queue length obtained by method P has minimum absolute error. The mean queue length obtained by method A is very close to that by method P, since the difference is only a first order term. The mean queue length obtained by method B is less accurate. In fact, it is very close to the Kingman's upper bound on the mean queue length under heavy traffic condition. The relative error under method P will approximately be $2/9(1-\rho) - 7/81(1-\rho)^2$, i.e., 2.1% for $\rho = 0.9$ and 4.1% for $\rho = 0.8$, as can be checked with Table 3.2. Method A has similar performance. The relative error of method B will be approximately $8/9(1-\rho) - 14/81(1-\rho)^2$, i.e., 8.7%

for $\rho = 0.9$ and 17.1% for $\rho = 0.8$, as can be checked with Table 3.2. Some numerical comparisons of methods P, A, and B are presented in Table 3.2 to check the correctness of the asymptotic analysis on mean queue lengths and their errors under various diffusion techniques for the E_2/M/1 system.

Table 3.2

MEAN QUEUE LENGTH FOR $E_2/M/1$ SYSTEM

ρ	Exact Result	Method P	Method B	Method A
0.95	14.331	14.487	14.962	14.492
0.90	6.829	6.974	7.425	6.985
0.85	4.327	4.463	4.887	4.477
0.80	3.075	3.200	3.600	3.219
0.75	2.323	2.438	2.813	2.460
0.70	1.820	1.925	2.275	1.95

4. ACCURACY ANALYSIS OF DIFFUSION APPROXIMATION BY SIMULATION AND NUMERICAL TECHNIQUES

In the previous section, we analyze the accuracy of mean queue lengths obtained by various diffusion approximation techniques for the M/G/1 system and the E_2/M/1 system where analytic solution of mean queue length is available. Now we continue the accuracy analysis on mean queue lengths obtained by various diffusion approximation techniques for those queueing systems where closed form expression for mean queue length is not available. Either simulation or numerical technique is employed to obtain an estimation of the mean queue length, depending upon which way is more convenient. Since simulation is only a statistical experiment and its convergence is slow under heavy traffic condition, not only the point estimation but also the 95% interval estimations are included to give a better feeling on the accuracy or convergence of the simulation.

In this section, we will concentrate on the cases where the coefficient of variation of the arrival process is less than 1. When the coefficient of variation of the arrival process exceeds 1, certain anomalies might happen, as we shall see in the next section. We will choose Erlang distribution to represent the interarrival arrival time distribution since its coefficient of variation is less than one. To be more specific, the squared coefficient of variation of an n stage Erlang distribution is equal to 1/n [9]. And we will use Erlang distribution and hyperexponential distribution to represent the service time distribution with coefficient of variation less than and greater than 1, respectively. A two stage hyperexponential density function has the following form:

$$\frac{\omega}{M_1} e^{-\frac{X}{M_1}} + \frac{(1 - \omega)}{M_2} e^{-\frac{X}{M_2}}$$

where

$$0 < \omega < 1$$

Apparently, it is the combination of two exponential distributions with mean M_1 and M_2, respectively. The probability of taking the first branch is ω, and that of taking the second branch is $1 - \omega$. Let us assume $M_2 < M_1$. After simple manipulation, we can express ω and M_1 in terms of M_2, M, and C, where M and C are the mean and squared coefficient of variation of the distribution function, respectively.

$$\omega = \frac{(M - M_2)^2}{M_2^2 - 2M_2M + \frac{M^2}{2}(C + 1)} \tag{4.1}$$

$$M_1 = \frac{M - (1 - \omega) M_2}{\omega} \tag{4.2}$$

Furthermore, for any M and C larger than 1, we can choose M_2 arbitrarily except that it must be in between 0 and M. The latter constraint will guarantee that the ω obtained from (4.1) will lie in between 0 and 1, and the M_1 obtained from (4.2) will be positive. Although various combinations of ω, M_1, and M_2 will lead to the same mean and variance, the higher moments of the distributions can be quite different. Since C can be chosen arbitrarily, we can get any value of coefficient of variation larger than one by using two stage hyperexponential distribution functions.

As pointed out earlier, all the results from simulations are expressed in terms of 95% confident interval estimations. For the $E_n/E_r/1$ system, the widths of the confident interval are less than 4% of the point estimations. For the $E_n/H_2/1$ system, the width of the confidence interval grows as the coefficient of variation and the traffic intensity increases. Nevertheless, even in the worst situation, when $C_s = 128$ and $\rho = 0.80$ in Table 4.5 of the $E_2/H_2/1$ system, the interval estimation of mean queue length is 203.7 ± 21 and the mean queue lengths obtained by methods P, B, and A are 206.4, 257.6, and 257.2, respectively. The superiority of method P is apparent in this case. That is to say, in all the cases where simulations are used, the simulation results are accurate enough to distinguish the relative performance of various diffusion approximations. Otherwise, numerical technique will be used.

We first examine the $E_n/E_r/1$ system. Table 4.1 compares the results when $\rho = 0.85$, $C_s = 0.5$, and $C_a = 1/2, 1/3, 1/4$ and $1/5$. Method A leads to the most accurate result within 3% relative error. Method B is inferior to the other techniques and can have a relative error at least up to 17%. Method P always has an error less than one half of the error in method B. Table 4.2 contains the results when $\rho = 0.8$, $C_s = 0.5$, and $C_a = 1/2, 1/3, 1/4, 1/5$. Similar patterns are again observed. As we shall see, the case where both coefficients of variation of interarrival time distribution and service time distribution are small, i.e., the $E_n/E_r/1$ system where both coefficients of variation are less than or equal to 0.5, is the only case where method P does not yield the best approximation.

Let us now examine the $E_r/H_2/1$ system. This is one of the cases where method P is much supreior to methods A and B. Table 4.3 contains the results when $\rho = 0.85$, $C_a = 0.5$, $C_s = 2, 4, 8, 16, 32$, and 64. The relative error of approximate mean queue length under method P is always very small in all the test cases. Methods A and B have very similar performance, and their relative errors can be at least up to 20% when $\rho = 0.85$. If we look at the table more carefully, we might find that the absolute errors in the approximate mean queue length under methods A and B are very close to $(\rho/2)C_s$. These are exactly the asymptotic absolute errors of approximate mean queue lengths under methods A and B in the M/G/1 system. Again, we observe the robustness of method P when the coefficient of variation of service time distribution is large. Table 4.4 compares the results for $\rho = 0.85$, $C_a = 1/3$, $C_s = 2, 4, 8, 16, 32$, and 64, and Tables 4.5 and 4.6 compare the results for $C_a = 0.5$, $C_s = 2, 4, 8, 16, 64$, and 128 when $\rho = 0.80$ and $\rho = 0.75$, respectively. Similar pattern is again observed.

Finally, we use numerical techniques to study the GI/M/1 system, namely the $E_3/M/1$ system and the D/M/1 system. The $E_2/M/1$ system has already been analyzed in Section 3, and we pointed out in (3.11) that the mean queue length of the GI/M/1 system is $\rho/(1-\sigma)$ where σ is the solution of the equation $A^*(\mu - \mu\sigma) = \sigma$. For the $E_2/M/1$ system, this equation is a second order equation, and we have a closed form for the root. But, for the $E_3/M/1$ and D/M/1 systems, the equations are third order and transcendental equations, respectively; so we have to

Table 4.1

MEAN QUEUE LENGTH FOR $E_n/E_r/1$ SYSTEM WHEN $\rho = 0.85$

C_a	C_s	Simulation	Method P	Method B	Method A
1/2	1/2	3.140 ± 0.039	3.258	3.471	3.069
1/3	1/2	2.694 ± 0.033	2.857	3.069	2.672
1/4	1/2	2.478 ± 0.024	2.656	2.869	2.473
1/5	1/2	2.347 ± 0.024	2.536	2.748	2.355

Table 4.2

MEAN QUEUE LENGTH FOR $E_n/E_r/1$ SYSTEM WHEN $\rho = 0.80$

C_a	C_s	Simulation	Method P	Method B	Method A
1/2	1/2	2.295 ± 0.022	2.400	2.600	2.230
1/3	1/2	1.986 ± 0.016	2.133	2.333	1.968
1/4	1/2	1.840 ± 0.013	2.000	2.200	1.838
1/5	1/2	1.750 ± 0.013	1.920	2.120	1.760

Table 4.3

MEAN QUEUE LENGTH FOR $E_2/H_2/1$ SYSTEM WHEN $\rho = 0.85$

C_s	M/M_2	Simulation	Method P	Method B	Method A
2	2	6.74 ± 0.19	6.87	7.72	7.3
4	3	11.52 ± 0.43	11.69	13.39	12.97
8	5	20.77 ± 0.87	21.32	24.72	24.30
16	9	40.63 ± 2.54	40.59	47.39	46.46
32	17	79.63 ± 6.46	79.12	92.72	92.30
64	33	152.3 ± 14	156.2	183.4	183.0

Table 4.4

MEAN QUEUE LENGTH FOR $E_3/H_2/1$ SYSTEM WHEN $\rho = 0.85$

C_s	M/M_2	Simulation	Method P	Method B	Method A
2	2	6.26 ± 0.14	6.47	7.32	6.90
4	3	10.97 ± 0.35	11.29	12.99	12.57
8	5	20.25 ± 0.97	20.92	24.32	23.90
16	9	40.20 ± 3.02	40.19	46.99	46.56
32	17	78.8 ± 6.9	78.72	92.32	91.89
64	33	152.0 ± 14	155.8	183.0	182.6

Table 4.5

MEAN QUEUE LENGTH FOR $E_2/H_2/1$ SYSTEM WHEN $\rho = 0.80$

C_s	M/M_2	Simulation	Method P	Method B	Method A
2	2	4.67 ± 0.09	4.80	5.60	5.21
4	3	7.83 ± 0.22	8.00	9.60	9.21
8	5	14.11 ± 0.53	14.40	17.60	17.20
16	9	27.24 ± 1.39	27.20	33.60	33.20
32	17	52.95 ± 3.02	52.8	65.6	65.2
64	33	102.4 ± 8	104.0	129.6	129.2
128	65	203.7 ± 21	206.4	257.6	257.2

Table 4.6

MEAN QUEUE LENGTH FOR $E_2/H_2/1$ SYSTEM WHEN $\rho = 0.75$

C_s	M/M_2	Simulation	Method P	Method B	Method A
2	2	3.44 ± 0.05	3.56	4.31	3.95
4	3	5.67 ± 0.12	5.81	7.31	6.94
8	5	10.08 ± 0.32	10.31	13.31	12.94
16	9	19.27 ± 0.83	19.31	25.31	24.94
32	17	37.39 ± 1.92	37.31	49.31	48.94
64	33	73.02 ± 4.73	73.31	97.31	96.94
128	65	146 ± 14	145.3	193.3	192.9

use numerical techniques to find the root. After simple manipulations, we get the following equations:

$$\sigma^3 - (2 + 9\rho)\,\sigma^2 + (1 + 9\rho + 27\rho^2)\,\sigma - 27\rho^3 = 0 \qquad \text{for the } E_3/M/1 \text{ system}$$

and

$$e^{-(1-\sigma)\mu} - \sigma = 0 \qquad \text{for the D/M/1 system}$$

In Tables 4.7 and 4.8, we compare the mean queue lengths obtained by numerical technique with those by various diffusion approximations for the $E_3/M/1$ and D/M/1 systems, respectively. Again, method P is the more accurate approximation method. The mean queue length obtained by method A is very close to that by method P. The relative error in method B can exceed those in methods A and B by 25% in some cases. Recall similar phenomenon appeared in our analysis on the $E_2/M/1$ system in Section 3.

Table 4.7

MEAN QUEUE LENGTH FOR E_3/M/1 SYSTEM

ρ	Exact Result	Method P	Method B	Method A
0.95	12.775	12.983	13.458	12.989
0.90	6.106	6.300	6.750	6.312
0.85	3.881	4.061	4.486	4.078
0.80	2.768	2.933	3.333	2.954
0.75	2.098	2.250	2.625	2.275
0.70	1.650	1.789	2.139	1.817

Table 4.8

MEAN QUEUE LENGTH FOR D/M/1 SYSTEM

ρ	Exact Result	Method P	Method B	Method A
0.95	9.664	9.975	10.450	9.983
0.90	4.661	4.950	5.400	4.965
0.85	2.991	3.258	3.683	3.280
0.80	2.154	2.400	2.800	2.427
0.75	1.651	1.875	2.250	1.906
0.70	1.313	1.517	1.867	1.551

5. ANOMALY WHEN THE COEFFICIENT OF VARIATION OF INTERARRIVAL TIME IS LARGER THAN ONE

In this section, we investigate the mean queue length of the queueing system where the coefficient of variation of the interarrival time is larger than 1. We will use the $H_2/M/1$ system as an example to demonstrate the anomaly since this is the case where analytic result is available. In the M/G/1 system, the mean queue length is given by the Pollaczek-Khinchin formula (3.7), and it depends only on the mean and variance of the service time distribution and the mean of the interarrival time distribution. This seems to be a support of the robustness of the diffusion approximation which only utilize the means and variances of the interarrival time and service time distributions to fit the parameters α and β of the Fokker-Plank equation and neglects the effect of higher moments of the distributions. However, after examining the $H_2/M/1$ system, we will find that higher moments of the interarrival time distribution do have a drastic effect on the mean queue length as the traffic intensity, ρ, deviates from 1. Since the two-stage hyperexponential distribution function is usually used when the distribution function is required to have high coefficient of variation, and is often encountered in computer system modelling, we will further analyze the regularity conditions on hyperexponential distributions for the diffusion approximation to be applicable. That is to say, we are interested in identifying the ranges of the parameters of the two-stage hyperexponential distribution when being used as the interarrival time distribution such that the performance of the queueing system, e.g., the mean queue length, can be estimated by the diffusion approximation accurately. It should not be too surprising that the ranges will shrink as ρ decreases or C_a increases. Recall the form of a two-stage hyperexponential distribution is $(\omega/M_1)\, e^{-X/M_1} + ((1-\omega)/M_2)\, e^{-X/M_2}$, where $M_2 < M_1$. The relations among the parameters are given in (4.1) and (4.2) with $M = 1/\lambda$.

From Table 5.1d, we observe that by choosing different ω, M_1, and M_2, the mean queue length varies over a wide range from 624.62 to 23.48 when $\rho = 0.95$, $C_a = 64$. Let us take a closer look at ω, M_1, and M_2 at two extreme cases:

Case 1:	$M_1 = 32.818$	$M_2 = 0.01$	$\omega = 3.018 \times 10^{-2}$
Case 2:	$M_1 = 3151.0$	$M_2 = 0.99$	$\omega = 3.175 \times 10^{-6}$

In both cases, the mean of interarrival time is 1, and the variance or the square coefficient of variation is 64. For the interarrival time distribution in Case 2, as we can see, M_2 is so close to the mean interarrival time, the first exponential density, $\omega/M_1\, e^{-X/M_1}$, has almost no effects on the mean interarrival time and only affects the variance. This is the reason why the mean queue length under the interarrival time distribution in Case 2 is almost equal to the mean queue length in M/M/1 which is equal to 19 when $\rho = 0.95$. That is to say, although having C_a equal to 64, the $H_2/M/1$ system in Case 2 behaves very much like an M/M/1 system. The interarrival time is generated according to the first exponential density $\omega/M_1\, e^{-X/M_1}$ so infrequently that we may ignore it when evaluating the performance of the queueing system. But in Case 1, M_2 is so close to zero, the first exponential density not only affects the variance of the interarrival time but also its mean. This is the reason why the mean queue length becomes so high, 624.62. The important fact to realize is that a large coefficient of variation does not necessarily mean that we have a lot of short interarrival times, and the fluctuation is quite high, as in Case 1; it may also mean there is few very long interarrival time which makes the coefficient of variation large without having serious effect on the fluctuation of the system.

Let us take a look at the mean queue length predicted by the diffusion approximations. It is around 587 by all three methods. This is quite close to the result under the interarrival time distribution in Case 1. This is not a surprise since mean queue length calculated by diffusion approximation is fairly close to the Kingman's upper bound on mean queue length as pointed out in Section 3. What we are interested in is, can we decide the applicability of diffusion approximation to the queueing system by just examining the parameters of the distribution. Surely, the applicable range will be affected by ρ and C_a. In Tables 5.1a through 5.1d, the mean queue lengths of the $H_2/M/1$ systems are tabulated for various combinations of M_1/M, M_2/M, ω which lead to the

Table 5.1a

MEAN QUEUE LENGTH WHEN $C_a = 2$

M_1/M	M_2/M	ω	Mean Queue Length	
			$\rho = 0.95$	$\rho = 0.85$
1.5051	0.01	6.622×10^{-1}	28.53	8.50
1.5263	0.05	6.435×10^{-1}	28.48	8.48
1.5556	0.1	6.183×10^{-1}	28.45	8.45
1.6250	0.2	5.614×10^{-1}	28.39	8.39
1.7143	0.3	4.949×10^{-1}	28.33	8.34
1.8333	0.4	4.186×10^{-1}	28.26	8.27
2.0000	0.5	3.333×10^{-1}	28.17	8.19
2.2500	0.6	2.424×10^{-1}	28.06	8.10
2.6667	0.7	1.525×10^{-1}	27.90	7.96
3.0000	0.75	1.111×10^{-1}	27.79	7.87
3.5000	0.8	7.407×10^{-2}	27.63	7.47
6.0000	0.9	1.961×10^{-2}	26.91	7.29
11.000	0.95	4.975×10^{-3}	25.76	6.78
50.998	0.99	2.000×10^{-4}	21.87	5.97

Table 5.1b

MEAN QUEUE LENGTH WHEN $C_a = 8$

M_1/M	M_2/M	ω	Mean Queue Length	
			$\rho = 0.95$	$\rho = 0.85$
4.5354	0.01	2.188×10^{-1}	85.55	25.47
4.6842	0.05	2.050×10^{-1}	85.35	25.32
4.8889	0.1	1.880×10^{-1}	85.12	25.12
5.3750	0.2	1.546×10^{-1}	84.67	24.67
6.0000	0.3	1.228×10^{-1}	84.11	24.12
6.8333	0.4	9.326×10^{-2}	83.39	23.43
8.0000	0.5	6.667×10^{-2}	82.42	22.51
9.7500	0.6	4.372×10^{-2}	81.02	21.21
12.667	0.7	2.507×10^{-2}	78.79	19.25
15.000	0.75	1.754×10^{-2}	77.03	17.83
18.500	0.8	1.130×10^{-2}	74.49	15.96
36.000	0.9	2.850×10^{-3}	62.89	10.53
71.000	0.95	7.138×10^{-4}	45.94	7.70
351.00	0.99	2.857×10^{-5}	23.16	6.00

Table 5.1c

MEAN QUEUE LENGTH WHEN $C_a = 32$

M_1/M	M_2/M	ω	Mean Queue Length	
			$\rho = 0.95$	$\rho = 0.85$
16.6565	0.01	5.947×10^{-2}	313.65	93.37
17.3158	0.05	5.502×10^{-2}	312.76	92.70
18.2222	0.1	4.966×10^{-2}	311.81	91.79
20.3750	0.2	3.965×10^{-2}	309.69	89.68
23.1429	0.3	3.064×10^{-2}	306.98	87.01
26.8333	0.4	2.270×10^{-2}	303.47	83.51
32.0000	0.5	1.587×10^{-2}	298.61	78.65
39.7500	0.6	1.022×10^{-2}	291.18	71.55
52.6667	0.7	5.773×10^{-3}	279.13	60.11
63.0000	0.75	4.016×10^{-3}	269.58	51.45
78.5000	0.8	2.574×10^{-3}	255.36	39.67
156.000	0.9	6.447×10^{-4}	187.90	13.23
311.000	0.95	1.613×10^{-4}	84.62	7.98
1551.00	0.99	6.452×10^{-6}	23.43	6.01

Table 5.1d

MEAN QUEUE LENGTH WHEN $C_a = 64$

M_1/M	M_2/M	ω	Mean Queue Length	
			$\rho = 0.95$	$\rho = 0.85$
32.818	0.01	3.018×10^{-2}	624.62	184.27
34.158	0.05	2.785×10^{-2}	616.40	182.61
36.000	0.1	2.507×10^{-2}	614.69	180.73
40.375	0.2	1.991×10^{-2}	610.10	176.40
46.000	0.3	1.532×10^{-2}	604.62	170.85
53.500	0.4	1.113×10^{-2}	596.79	163.54
64.000	0.5	7.874×10^{-3}	586.57	153.35
79.750	0.6	5.054×10^{-3}	571.47	138.28
106.00	0.7	2.850×10^{-3}	545.90	113.66
127.00	0.75	1.980×10^{-3}	525.69	94.63
158.50	0.8	1.269×10^{-3}	497.30	67.99
316.00	0.9	3.174×10^{-4}	347.83	14.10
631.00	0.95	7.936×10^{-5}	116.30	8.02
3151.0	0.99	3.175×10^{-6}	23.48	6.01

same means and squared coefficients of variation, which are 2, 8, 32, and 64, respectively, when $\rho = 0.95$ and 0.85. M_1/M and M_2/M can be viewed as the normalized means of the first exponential and second exponential branches, respectively. As we can observe from Table 5.1, as M_2/M increases, the mean queue length decreases. The mean queue length drops sharply as M_2/M approaches 1. The larger the coefficient of variation of the interarrival time is, the larger the variation of mean queue length can be. Nevertheless, the mean queue length does not vary too much over a substantial range of the value of M_2/M. This gives us some hope that diffusion approximation may be applied in this case. In Tables 5.2a and 5.2b, the mean queue lengths of the $H_2/M/1$ systems when C_a = 2, 4, 8, 16, 32, and 64 are tabulated for $\rho = 0.95$ and 0.85, respectively. The exact mean queue lengths for $M_2/M = 0.2$, 0.5, and 0.7 are also included in Table 5.2a. Similarly, those for $M_2/M = 0.1$, 0.4, and 0.6 are also included in Table 5.2b. As we can see, various diffusion approximations lead to similar results and they provide reasonable approximations to the analytic results under those M_2/M ratios. Combining the results from Tables 5.1 and 5.2, we can conclude that, for ρ = 0.95, the range of M_2/M where diffusion approximation can be applied is

$$\frac{M_2}{M} \leq 0.75$$

within 15% accuracy for $C_a \leq 64$. This is a very conserved bound. When C_a is close to 1, the applicable range is actually larger than the specified range. As we can see from Table 5.1a, for the case $C_a = 2$ and $\rho = 0.95$, even when M_2/M increases to 0.95, the variation of mean queue length is not substantial, and the mean queue length from diffusion approximation (see Table 5.2a) is acceptable. As the traffic intensity, ρ, decreases, the applicable range shrinks very quickly, as can also be observed from Table 5.1. In the case $\rho = 0.85$, the applicable range is around

$$\frac{M_2}{M} \leq 0.6 \text{ to } 0.65$$

with 15% accuracy for $C_a \leq 64$.

Table 5.2a

MEAN QUEUE LENGTH FOR $H_2/M/1$ SYSTEM WHEN $\rho = 0.95$

C_a	Method P	Method B	Method A	Analytic		
				$M_2/M = 0.2$	$M_2/M = 0.5$	$M_2/M = 0.7$
2	28.0	28.5	28.0	28.4	28.2	27.9
4	46.1	46.5	46.1	47.2	46.3	45.1
8	82.2	82.6	82.2	85.1	82.4	78.8
16	154.4	154.8	154.4	160.7	154.5	145.7
32	298.8	299.2	298.8	309.7	298.6	279.1
64	587.6	588	587.6	610.1	586.8	544.9

Table 5.2b

MEAN QUEUE LENGTH FOR $H_2/M/1$ SYSTEM WHEN $\rho = 0.85$

C_a	Method P	Method B	Method A	Analytic		
				$M_2/M = 0.1$	$M_2/M = 0.4$	$M_2/M = 0.6$
2	8.08	8.50	8.08	8.45	8.27	8.10
4	12.9	13.3	12.9	14.0	13.4	12.6
8	22.5	23.0	22.5	25.1	23.4	21.2
16	41.8	42.1	41.8	47.4	43.5	38.1
32	80.3	80.8	80.3	91.8	83.5	71.5
64	157.4	157.8	157.4	180.7	163.5	138.3

The mean queue lengths for more general service time distributions under H_2 input are not available analytically. It is hard to do a thorough investigation of the applicable range of diffusion approximation in this case since the exact mean queue length can only be estimated through long simulations under heavy traffic condition or tedious numerical techniques. Nevertheless, we select certain combinations of M_1, M_2, ω which fall inside the applicable range of diffusion approximation to $H_2/M/1$ systems and simulate the mean queue lengths for several $H_2/H_2/1$ and $H_2/E_r/1$ systems. In Table 5.3, the mean queue lengths obtained by various diffusion approximations are compared with that obtained by simulations for the $H_2/E_2/1$ system when $\rho = 0.85$ and $C_a = 64, 32, 16, 8$, and 4. The second column $M_2\lambda$ specifies the extra degree of freedom in the interarrival time distribution and also is the criterion we use to test the applicability of diffusion approximation in the $H_2/M/1$ system. The simulation results seem to be quite close to the results obtained by various diffusion approximations. All three diffusion approximation techniques yield very similar results. In Table 5.4, the mean queue lengths obtained by various diffusion approximations are compared with those obtained by simulations for the $H_2/H_2/1$ system when $\rho = 0.85$ under various combinations of C_a and C_s. The interarrival time distribution is assumed to have the same form as before. The service time distribution is also assumed to have a similar form as $(\omega/M_{s1})\, e^{-X/M_{s1}} + ((1 - \omega)/M_{s2})\, e^{-X/M_{s2}}$ with mean $1/\mu$. The second column is the same as that in Table 5.3. The fourth column is only used to specify the extra degree of freedom in the service time distribution. The results under various diffusion approximations are again quite acceptable. Furthermore, methods B and A yield very similar results, and the result from method P is somewhat smaller. Recalling the asymptotic expressions for mean queue lengths in Section 2, we know that the mean queue length obtained by method P is smaller than those obtained by methods A and B by $(\rho/2)\, C_s$. This is indeed the case as can be checked from Table 5.4. Since the results obtained by methods A and B are closer to the Kingman's upper bound, the mean queue length obtained by method P seems to cover a wider range of the parameters of the distribution within $\pm 15\%$ accuracy. That is to say, method P seems to be preferable unless $M_2\lambda$ is very close to 0 for the $H_2/H_2/1$ system.

Table 5.3

MEAN QUEUE LENGTH FOR $H_2/E_2/1$ SYSTEM WHEN $\rho = 0.85$

C_a	λM_2	Simulation	Method P	Method B	Method A
64	0.2	163 ± 16	156.2	156.4	156.0
32	0.2	83.5 ± 5.5	79.1	79.3	78.9
32	0.5	71.9 ± 6.3	79.1	79.3	78.9
16	0.15	43.5 ± 3.1	40.6	40.8	40.4
16	0.5	39.5 ± 2.6	40.6	40.8	40.4
8	0.3	22.6 ± 1.1	21.3	21.5	21.1
8	0.45	21.5 ± 0.8	21.3	21.5	21.1
4	0.3	12.2 ± 0.4	11.7	11.9	11.5
4	0.5	11.5 ± 0.4	11.7	11.9	11.5

Table 5.4

MEAN QUEUE LENGTH FOR $H_2/H_2/1$ SYSTEM WHEN $\rho = 0.85$

C_a	λM_2	C_s	μM_{s2}	Simulation	Method P	Method B	Method A
64	0.2	64	0.2	312 ± 33	309	336	336
64	0.2	32	0.5	237 ± 19	232	246	245
32	0.2	64	0.6	230 ± 20	232	259	259
16	0.3	16	0.3	74.8 ± 5.0	77.9	84.7	84.3
8	0.75	4	0.75	25.0 ± 1.4	29.8	31.5	31.0
4	0.6	4	0.6	18.4 ± 1.0	20.1	21.8	21.4
4	0.75	4	0.75	18.5 ± 0.6	20.1	21.8	21.4

From then on, we will use the term "type A" hyperexponential distribution to denote the hyperexponential distribution having the property that M_2/M is not "close" to 1. We will be vague on the exact range of the parameters that type A hyperexponential distributions must satisfy, i.e., we only define it in a qualitative way, not in a quantitative way. For more complicated queueing systems, the exact range of the parameters of type A hyperexponential distributions such that the diffusion approximation can be applied to obtain a reasonable estimate of mean queue length, utilization, etc. may be different, depending on the network topology and traffic intensity at the server. But, as long as M_2/M is not far from zero, the approximation should be applicable. Furthermore, we will use the term "type B" hyperexponential distribution to represent the other extreme, i.e., M_2/M is close to 1.

Although hyperexponential distributions have been used throughout the section as interarrival time distributions, we expect the results should hold for other general distributions. That is to say, as long as the high variation of interarrival time is due to a large number of short interarrival times instead of a few very long interarrival times, diffusion approximation should be applicable. Similarly, the terms type A and type B distributions are used to denote the two types of distributions with high coefficients of variation, respectively.

6. CLOSED TWO SERVER SYSTEMS (CPU/DTU MODEL)

After completing the discussion on diffusion approximation to single server systems, we now consider applying diffusion approximation to closed two server systems, as shown in Fig. 6.1. In computer system modelling, the closed two server system is often used to represent the CPU (central processing unit) and DTU (data transfer unit) operating under fixed degree of multiprogramming. From then on, we will call the two servers CPU and DTU, respectively. This model has been analyzed by Gaver and Shedler [2] using diffusion approximation with reflecting boundaries and usual way to estimate α and β when the CPU service time which is the time between page faults under this particular interpretation has exponential distribution, i.e., when its coefficient of variation is equal to 1. Later on, Gaver and Shedler [3] use Wald's identity to fit the ratio of $2\beta/\alpha$ and analyze the same system when the CPU service time has hyperexponential distribution, i.e., when its coefficient of variation is larger than 1. Gelenbe [5] also analyzed this system using the Feller's elementary return process and the usual way to estimate diffusion parameters.

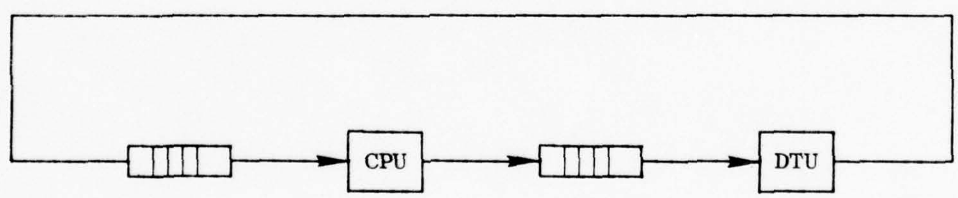


Fig. 6.1. CLOSED TWO SERVER SYSTEM (CPU-DTU SYSTEM).

The anomalies in Section 5 lead us to suspect the same kind of problems may exist in the two server closed queueing networks if one or more servers have hyperexponential service time distributions. This is simply because the departure process of any server is the arrival process of the other server. Before we proceed to investigate the anomalies, let us first examine the various diffusion approximation techniques on the two server closed queueing network just mentioned in some detail. We will concentrate on the CPU system and try to estimate the utilization or the mean queue length of the CPU system. This measure is of practical importance since it can be used as an indicator of the stationary system performance. Assume the fixed degree of multiprogramming is M. Let $P(X)$ be the stationary probability density function for the diffusion process in $(0,M)$, where X is the number of programs in the CPU queue and $M-X$ is the number of programs in the DTU queue. Furthermore, let us assume the mean service time of the CPU and DTU are $1/\mu$ and $1/\lambda$, and the variance of the CPU and DTU are σ_s^2 and σ_a^2, respectively.

The first diffusion approximation technique which we are going to examine is the method proposed by Gaver and Shedler [2]. In this method, we have

$$\alpha = \sigma_a^2\lambda^3 + \sigma_s^2\mu^3$$

$$\beta = \mu - \lambda$$

as usual. Let $F(X)$ be the stationary distribution function of the queue length at CPU. Imposing a reflection boundary at $X = 0$ and normalizing the probability mass between 0 and M to one, we get

$$F(X) = \frac{1 - A\, e^{rX}}{1 - A\, e^{rM}}$$

where

$$r = \frac{2\beta}{\alpha} = \frac{2(\mu - \lambda)}{\sigma_a^2\lambda^3 + \sigma_s^2\mu^3}$$

by solving the Fokker-Plank equation (2.1). The unknown constant A is chosen such that

$$\lim_{M\to\infty} F(0) = 1 - \rho$$

After simple manipulation, we get

$$A = \rho$$

Hence,

$$F(X) = \frac{1 - \rho\, e^{rX}}{1 - \rho\, e^{rM}}$$

We will refer to this method as method G1 later on.

The second diffusion approximation technique is again due to Gaver and Shedler [3] to handle the case where the coefficient of variation of the CPU service time is larger than one. Let $G(S)$ be the Laplace-Stieltjes transform of the CPU service time distribution and $H(S)$ be the Laplace-Stieltjes transform of the DTU service time distribution. Furthermore, let S^* be the positive solution of

$$G(-S)\, H(S) = 1$$

By Wald's identity, we get

$$r = \frac{2\beta}{\alpha} = \ln G(S^*)$$

Following the same argument as the previous method, we get

$$F(X) = \frac{1 - A\, e^{rX}}{1 - A\, e^{rM}}$$

Now, using the fact that the long run input rate to the CPU, $\lambda F(M-1)$, must equal to the long run output rate from the CPU, $\mu(1 - F(0))$, we get

$$A = \frac{\rho}{1 + \rho\, e^{-r(M-1)} - e^{-rM}}$$

We will refer to this method as method G later on.

Finally, we consider the method proposed in Gelenbe [5] with a slightly different argument from [5] making it coherent with Section 2. Again, we refer to this method as method B.

Now there are two boundaries: 0 and M. When the number of programs in the CPU queue is M, the DTU is idle and no new arrival will occur. It is assumed that the CPU queue length will jump to M-1 after an exponential holding time with mean h_M. This is similar to the way the boundary at $X = 0$ is handled in Section 2. The boundary at $X=0$ is still handled in the same way as in Section 2, where the mean holding time is assumed to be h_0. As noted earlier, restricting the holding time distribution on the boundaries to be exponentially distributed does not mean to impose any restriction on the service time distribution of the CPU or DTU. It is just a conceptual help for us to handle the boundary conditions. Now we are facing a more serious problem than we encountered in the GI/G/1 system. Here, not only the two mean holding times at the boundary $X = 0$ and $X = M$ are unknown, but also the probabilities of having empty queue and full queue are unknown. As we shall see, the boundary equations can only be used to solve two unknowns, i.e., we must find approximate values for two of the four unknown parameters. Since the probability of empty queue is directly related to the CPU utilization, a quantity of major concern, we will try to find reasonable estimations for h_1 and h_M, the mean holding times at the boundaries, and hope that the errors in these estimations will have minor influence on other quantities of interest. Recall the GI/G/1 system in Section 2 where the approximate holding time at $X = 0$ obtained by the diffusion approximation is equal to $1/\lambda$, which is the mean interarrival time. Hence, we will set h_0 equal to $1/\lambda$, the mean service time of the DTU since, when the DTU is busy, the mean interarrival time to the CPU is equal to the mean service time of the DTU. By similar argument, we will set h_M equal to $1/\mu$, the mean service time of the CPU.

At steady state, we have the following equations:

$$\begin{cases} \frac{1}{2}\alpha\frac{\partial^2}{\partial X^2}P(X) - \beta\frac{\partial}{\partial X}P(X) = -\lambda M_1\delta(X-1) - \mu M_2\delta(X-M+1) \\ \lim\limits_{X\to 0}\frac{1}{2}\alpha\frac{\partial}{\partial X}P(X) - \beta P(X) = \lambda M_1 \\ \lim\limits_{X\to M}\frac{1}{2}\alpha\frac{\partial}{\partial X}P(X) - \beta P(X) = \mu M_2 \end{cases}$$

where $\delta(\cdot)$ is the Dirac density function and $\delta(X-1)$ and $\delta(X-M+1)$ represent the probability density function of the point from which the diffusion process starts once again immediately after a jump from the boundaries 0 and M, respectively; M_1 and M_2 are the probability masses concentrated on the lower boundary and upper boundary, respectively.

Furthermore, we have the following boundary conditions:

$$\lim_{X\to 0} P(X) = \lim_{X\to M} P(X) = 0$$

Solving the above equations, we get

$$P(X) = \begin{cases} \dfrac{\lambda M_1}{\beta}\,(1 - e^{-rX}) & 0 \leq X \leq 1 \\[2ex] \dfrac{\lambda M_1}{\beta}\,(e^{-r} - 1)\, e^{rX} & 1 \leq X \leq M - 1 \\[2ex] \dfrac{\lambda M_2}{\beta}\,(e^{r(X-M)} - 1) & M - 1 \leq X \leq M \end{cases} \tag{6.1}$$

where

$$M_2 = \frac{\lambda M_1}{\mu}\, e^{r(M-1)}$$

and

$$r = \frac{2\beta}{\alpha}$$

Also, using

$$\int_0^M P(X)\, dX + M_1 + M_2 = 1$$

we get

$$M_1 = (1 - \rho)\; 1 - \rho^2 \left(e^{r(M-1)}\right)^{-1} \tag{6.2}$$

A more detailed derivation of the result is given in Gelenbe [5]. We will further consider the problem of discretization of the probability density function in the neighborhood of integer valued point $x = i$ in order to approximate π_i, the stationary probability of having i jobs in the system. The following way of discretization is proposed.

$$\begin{cases} \pi_0 = M_1 \\ \pi_1 = \int_0^{3/2} P(X)\,dX \\ \pi_i = \int_{i-1/2}^{i+1/2} P(X)\,dX \qquad 2 \le i \le M-2 \\ \pi_{M-1} = \int_{M-3/2}^{M} P(X)\,dX \\ \pi_M = M_2 \end{cases}$$

After simplification, we get

$$\begin{cases} \pi_0 = (1-\rho)\left(1 - \rho^2 e^{r(M-1)}\right)^{-1} \\ \pi_1 = \dfrac{\pi_0 \rho}{(1-\rho)\,r}\left(r + 1 - e^r + \left(e^{\frac{3}{2}r} - e^r\right)\left(e^{-r} - 1\right)\right) \\ \pi_i = -\dfrac{\pi_0 \rho}{(1-\rho)\,r}\, e^{\left(i+\frac{1}{2}\right)r}\left(1 - e^{-r}\right)^2 \qquad \text{for } 2 \le i \le M-2 \\ \pi_{M-1} = \dfrac{\pi_0 \rho}{(1-\rho)\,r}\, e^{r(M-1)}\left(e^{-r/2} - e^{-3r/2} - r\right) \\ \pi_M = \rho\pi_0\, e^{r(M-1)} \end{cases} \tag{6.3}$$

In the G/G/1 system, we find that the accuracy of diffusion approximation can be improved by defining the diffusion parameter β as $\lambda C_a + \mu g C_s$ where g is set to ρ, the traffic intensity or the utilization of the server. For the closed two server system, the utilization of each server is not apparent. From experimental results, it seems to be that setting g equal to λ/μ for $C_s > 1$ and 1 otherwise may improve the accuracy. We will denote this method as method P.

After examining the various diffusion approximation methods, now let us consider the case where the coefficient of variation of the service time distribution at the CPU is large, as is often the case. If the service time distribution at the DTU is exponentially distributed, the system is analytically tractable since it can be viewed as a M/G/1 queueing system with finite waiting room [19]. We summarize the result for stationary mean queue length as follows.

$$\pi_i^M = K_M \pi_i \qquad \text{for} \quad 1 \leq i < M$$

$$\pi_M^M = 1 - \frac{1 - K_M(1 - \rho)}{\rho}$$

and

$$K_M = \frac{1}{1 - \rho\left(1 - \sum_{i=0}^{M-1} \pi_i\right)}$$

where

π_i^M is the stationary probability that the queue length is equal to i when the capacity of the waiting room is M

π_i is the stationary queue length distribution of the M/G/1 system

Hence, we will assume the service time at DTU has exponential distribution and use the analytic result to analyze the accuracy and applicability of diffusion approximation. The answer to the following three questions are of major interest: (1) Does the mean queue length or the utilization of the CPU vary if type B hyperexponential distribution instead of type A hyperexponential distribution is used for the service time distribution of CPU, (2) Does the diffusion approximation give a reasonable approximation to the mean queue length and utilization when type A hyperexponential distribution is used for service time distribution of CPU, (3) Does the service time of the CPU often have type A hyperexponential distribution, i.e., we are more interested in the applicability to computer system modelling.

The hyperexponential distribution of the CPU service time is assumed to have to form $(\omega/M_1)\ e^{-X/M_1} + ((1-\omega)/M_2)\ e^{-X/M_2}$ with mean $1/\mu$ where $M_2 < M_1$. From Tables 6.1 and 6.2, we see that the mean queue length and utilization change as the parameters of the hyperexponential distribution change. Similar phenomenon on CPU utilization in the M/G/1/N system is observed by Price [24]. As $M_2\mu$ decrease, i.e., the number of requests having short CPU interval increases, the analytic results become very close to the results obtained by both diffusion approximation methods.

Again, we see the diffusion approximation gives a good approximation of the performance under type A hyperexponential service time distribution. We also expect the results can be generalized to more general distributions. That is to say, diffusion approximation will be applicable if the large variation of service time is due to a lot of short service times. If it is due to a few long service times, diffusion approximation can only be used to obtain a lower bound of the performance.

The third question is hard to answer in general. In [3], three sets of data on the CPU utilization and the mean and variance of the CPU service time, which is the time between page faults, are given. The mean and variance of the CPU service time are gathered from actual program data. The results on CPU utilization are obtained by trace driven simulations of that queueing system. The DTU service time is assumed to be constant to account for the average access time along with the time to transfer a page of information. In Tables 6.3, 6.4, and 6.5, we compare the results obtained by three different diffusion approximation techniques, i.e., methods P, G, and B, with the analytic result obtained by semi-Markov analysis [22]. The parameter M_2's of the hyperexponential distributions taken in [3] are all less than four tenths of their means, respectively, so they should be type A hyperexponential distributions.

Table 6.1

MEAN QUEUE LENGTH OF CPU WHERE CPU HAS H_2 SERVICE TIME DISTRIBUTION WITH MEAN $1/\mu = 0.9$ AND SQUARED COEFFICIENTS OF VARIATION $C_S = 16$ AND DTU HAS EXPONENTIAL SERVICE TIME DISTRIBUTION WITH MEAN $1/\lambda = 1$

$M_2\mu$ / Number of Jobs	0.99	0.9	0.7	0.5	0.3	0.2	0.1	0.025	Diffusion Approximation	
									Method P	Method B
4	1.787	1.764	1.736	1.742	1.778	1.806	1.840	1.867	1.877	1.878
8	3.284	3.118	3.035	3.186	3.402	3.510	3.613	3.687	3.711	3.717
12	4.508	4.148	4.206	4.632	5.034	5.207	5.363	5.471	5.505	5.520
16	5.489	4.970	5.387	6.104	6.663	6.888	7.086	7.220	7.261	7.287
20	6.260	5.680	6.601	7.587	8.280	8.548	8.780	8.935	8.981	9.022
24	6.858	6.340	7.841	9.073	9.879	10.183	10.443	10.615	10.665	10.724
28	7.316	6.987	9.100	10.552	11.457	11.792	12.075	12.262	12.315	12.395
32	7.665	7.640	10.369	12.019	13.011	13.373	13.676	13.874	13.930	14.033
36	7.932	8.307	11.641	13.472	14.540	14.924	15.245	15.454	15.511	15.641
40	8.136	8.991	12.913	14.907	16.042	16.446	16.782	17.000	17.058	17.217

Table 6.2

UTILIZATION OF CPU WHERE CPU HAS H_2 SERVICE TIME DISTRIBUTION WITH MEAN $1/\mu = 0.9$ AND SQUARED COEFFICIENTS OF VARIATION $C_s = 16$ AND DTU HAS EXPONENTIAL SERVICE TIME DISTRIBUTION WITH MEAN $1/\lambda = 1$

$M_2\mu$ / Number of Jobs	0.99	0.9	0.7	0.5	0.3	0.2	0.1	0.025	Diffusion Approximation	
									Method P	Method B
4	0.7533	0.7306	0.6813	0.6355	0.5952	0.5776	0.5617	0.5508	0.5478	0.5418
8	0.8327	0.7970	0.7286	0.6807	0.6484	0.6359	0.6253	0.6182	0.6166	0.6069
12	0.8610	0.8184	0.7495	0.7106	0.6869	0.6781	0.6707	0.6658	0.6650	0.6538
16	0.8743	0.8282	0.7649	0.7342	0.7164	0.7100	0.7046	0.7011	0.7008	0.6892
20	0.8816	0.8339	0.7777	0.7533	0.7398	0.7350	0.7310	0.7284	0.7284	0.7168
24	0.8857	0.8379	0.7888	0.7692	0.7587	0.7550	0.7520	0.7501	0.7502	0.7389
28	0.8882	0.8412	0.7983	0.7825	0.7743	0.7715	0.7691	0.7677	0.7679	0.7570
32	0.8897	0.8440	0.8067	0.7938	0.7873	0.7851	0.7834	0.7822	0.7825	0.7721
36	0.8907	0.8466	0.8141	0.8036	0.7984	0.7967	0.7953	0.7945	0.7948	0.7848
40	0.8913	0.8490	0.8207	0.8121	0.8080	0.8066	0.8055	0.8049	0.8053	0.7957

Table 6.3

CPU UTILIZATION

$1/\mu = 17026, \quad \sigma_s^2 = 0.39780 \times 10^{10}, \quad M_2 = 3682, \quad 1/\lambda = 20{,}000$

Number of Jobs in the System	Semi-Markov	Method P	Method G	Method B
2	0.5316	0.4934	0.5993	0.4887
3	0.5548	0.5223	0.6667	0.5141
4	0.5752	0.5475	0.7064	0.5367
5	0.5935	0.5696	0.7326	0.5568
6	0.6098	0.5892	0.7511	0.5749
7	0.6245	0.6067	0.7650	0.5912
8	0.6379	0.6223	0.7757	0.6060
9	0.6500	0.6364	0.7842	0.6195
10	0.6611	0.6491	0.7911	0.6318

Table 6.4

CPU UTILIZATION

$1/\mu = 4871, \quad \sigma_s^2 = 0.26492 \times 10^{9}, \quad M_2 = 1929, \quad 1/\lambda = 20{,}000$

Number of Jobs in the System	Semi-Markov	Method P	Method G	Method B
2	0.2216	0.2169	0.2216	0.2022
3	0.2286	0.2285	0.2313	0.2077
4	0.2333	0.2350	0.2361	0.2124
5	0.2366	0.2387	0.2388	0.2165
6	0.2388	0.2408	0.2404	0.2201
7	0.2403	0.2420	0.2415	0.2231
8	0.2413	0.2426	0.2422	0.2258
9	0.2420	0.2430	0.2426	0.2281
10	0.2425	0.2432	0.2429	0.2301

Table 6.5

CPU UTILIZATION

$1/\mu = 10735, \quad \sigma_s^2 = 0.12313 \times 10^{10}, \quad M_2 = 2953, \quad 1/\lambda = 20{,}000$

Number of Jobs in the System	Semi-Markov	Method P	Method G	Method B
2	0.4076	0.3863	0.4249	0.3704
3	0.4281	0.4147	0.4579	0.3887
4	0.4449	0.4368	0.4764	0.4045
5	0.4587	0.4544	0.4882	0.4183
6	0.4702	0.4685	0.4964	0.4304
7	0.4798	0.4799	0.5024	0.4410
8	0.4879	0.4893	0.5070	0.4505
9	0.4948	0.4970	0.5106	0.4588
10	0.5006	0.5033	0.5136	0.4663

As expected, the diffusion approximations are very close to the analytic result, and method P seems to have better overall performance among the three diffusion approximations. In Tables 6.6a, 6.6b, and 6.6c, we compare all of them with the results obtained by trace driven simulations in [3], and the results are again very close. Hence, we can say, at least in that computer environment, the assumption of having type A hyperexponential service time distribution is very reasonable. Lewis and Shedler [20], based on a statistical analysis of actual computer program address traces, presented a semi-Markov model for the point process of page exceptions. Although the detailed stochastic structure of that model is more complicated, it does have the property that it consists of a large number of short interfault time similar to that of type A hyperexponential distribution.

For the case where the coefficient of variation of CPU is small (<1), methods P and B are equivalent. In [5], comparison of methods B and G1, with results obtained by semi-Markov analysis, shows that both methods are very accurate, and method B is a little bit better.

The model in Fig. 6.1 can also be generalized to include a self loop at each server, as shown in Fig. 6.2. To account for the effect of the self loop, we treat each server, including its self loop, as a single entity and consider the effect of the self loop as an internal interaction which is transparent to other parts of the system. That is to say, we will replace the server with a self loop by an equivalent* one without a self loop. The interdeparture time of the equivalent server is, in fact, the service completion time seen by DTU. The mean and variance of the service time of the equivalent CPU are $(1-\theta)/\mu$ and $(\sigma_s^2/(1-\theta) + \theta/(\mu^2(1-\theta)^2))$, respectively. The two quantities are derived below.

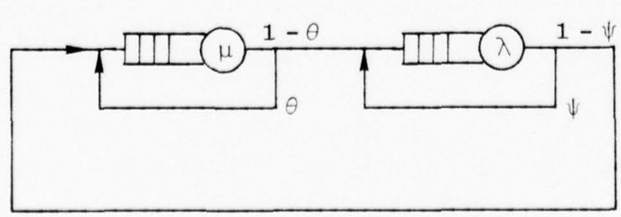


Fig. 6.2. CPU-DTU MODEL WITH SELF LOOPS.

* in the sense that queue length distribution and departure rate to the other server are preserved.

Table 6.6a
CPU UTILIZATION

$1/\mu = 17026, \quad \sigma_s^2 = 0.39780 \times 10^{10}, \quad 1/\lambda = 20,000$

Number of Jobs	Trace	Semi-Markov	Method P	Method G	Method B
3	0.538	0.5548	0.5223	0.6667	0.5141
6	0.546	0.6098	0.5892	0.7511	0.5749

Table 6.6b
CPU UTILIZATION

$1/\mu = 4871, \quad \sigma_s^2 = 0.26492 \times 10^{9}, \quad 1/\lambda = 20,000$

Number of Jobs	Trace	Semi-Markov	Method P	Method G	Method B
3	0.227	0.2286	0.2285	0.2313	0.2077
6	0.229	0.2388	0.2408	0.2404	0.2201

Table 6.6c
CPU UTILIZATION

$1/\mu = 10735, \quad \sigma_s^2 = 0.12313 \times 10^{10}, \quad 1/\lambda = 20,000$

Number of Jobs	Trace	Semi-Markov	Method P	Method G	Method B
3	0.419	0.4281	0.4147	0.4579	0.3887
6	0.425	0.4702	0.4685	0.4964	0.4304

Let N be the number of service completions at CPU in between jobs arriving at the DTU (including the last one). Then, N is a geometrically distributed random variable with mean $1/(1-\theta)$ and variance $\theta/(1-\theta)^2$. Let X be the random variable which represents the service time of CPU and Y be the random variable which represents the service time of the equivalent CPU without self loop. By assumption,

$$E\{X\} = \frac{1}{\mu} \quad \text{and} \quad Var\{X\} = \sigma_s^2$$

Clearly,

$$\begin{aligned} E\{Y\} &= E\{E\{Y|N\}\} \\ &= E\{NE\{X\}\} \\ &= \frac{1}{\mu(1-\theta)} \end{aligned}$$

Using the identity for conditional variance [14], we get

$$\begin{aligned} Var\{Y\} &= E\{Var\{Y|N\}\} + Var\{E\{Y|N\}\} \\ &= E\{N\,Var\{X\}\} + Var\{NE\{X\}\} \\ &= E\left\{N\sigma_s^2\right\} + Var\left\{\frac{N}{\mu}\right\} \\ &= \frac{\sigma_s^2}{1-\theta} + \frac{\theta}{\mu^2(1-\theta)^2} \end{aligned}$$

Similarly, we can derive the mean and variance of the service time of equivalent DTU without self loop. The mean and variance of the service time are equal to $(1-\psi)/\lambda$ and $(\sigma_a^2/(1-\psi) + \psi/(\mu^2(1-\psi)^2))$, respectively. After obtaining the means and variances of equivalent servers without self loop, we reduce the model to the original closed two-server queueing model.

When both stages have exponential service times, the model in Fig. 6.2 is equivalent to that in Fig. 6.1 with service rates $\mu(1-\theta)$ and $\lambda(1-\psi)$ at CPU and DTU, respectively. The forms of the service time distributions do not change in this case.

7. GENERAL QUEUEING NETWORKS

Finally, let us consider applying diffusion approximation to analyze the performance of queueing networks. Kobayashi [10] proposed that queueing processes of a general queueing network be approximated by a vector-valued diffusion process. The interactions among different queueing processes are explicitly considered in the diffusion equations in terms of the variance-covariance matrix. The joint queue length distribution is expressed in a product form of the marginal queue size distributions. This solution form suggests us to treat each queue separately by properly taking into account the interaction among different queues [12]. From our analytic and experimental data presented in the previous section, we know the accuracy of diffusion approximation is extremely good on single server system except for certain pathological cases given in Section 5. So, the success on decomposing queueing network into separate single server systems solely relies on whether we can find a good estimation of the coefficient of variation of the interarrival time distribution or the coefficient of variation of the interdeparture time distribution at each server such that the correlations among servers are not ignored after decomposition. Two different methods to estimate the coefficient of variation of the interdeparture time at each server have been proposed by Reiser and Kobayashi [12] and Gelenbe [6], respectively. The method proposed by Gelenbe tries to take into account the effect of idle period on the coefficient of variation of interdeparture time which is neglected by the other method and seems to lead to better results. Nevertheless, this method is more complicated in the sense that matrix inversion is involved. In this section, we propose a simpler way to estimate the coefficient of variation of the interdeparture time distribution than that by Gelenbe, yet the effect of idle period is taken into account. The values obtained by both methods are close to each other. Furthermore, all the demonstrating examples given by the previous authors to show the accuracy on decomposing queueing network into separate queues using diffusion approximation are under the condition that the coefficient of variation is not large, mainly less than or equal to 2. We present an anomaly which has been overlooked in the past, i.e., certain network topology can only be decomposed into separate single servers when all the service times and external interarrival times are "nearly" exponentially distributed. When the coefficients of variation of some of the service time or external interarrival time distributions deviates further from 1 or the traffic intensity decreases, the decomposition of this kind of queueing network will not be feasible if we still adopt the conventional way to estimate the coefficients of variation of interdeparture times or diffusion parameters.

Let us first examine the method in [12] proposed by Reiser and Kobayashi to estimate the coefficient of variation of the interarrival time. In their treatment, the coefficient of variation of the service time is taken to be the coefficient of variation of the interdeparture time as a simple approximation. Furthermore, the departure processes from different servers are treated as independent renewal processes. After considering the fact that the departure process of the i^{th} server are only active ρ_i percent of the time, where ρ_i is the utilization of the i^{th} server, they obtain the following expression for C_a^i, the squared coefficient of variation of the interarrival time of the i^{th} server:

$$C_a^i = \frac{1}{\lambda_i} \sum_{j=0}^{n} \left[(C_j - 1)\, P_{ji} + 1\right] \lambda_j P_{ji} \tag{7.1}$$

Furthermore, C_j, the squared coefficient of variation of the interdeparture time of the j^{th} server, is approximated by

$$C_j = C_s^j \tag{7.2}$$

where

λ_j is the arrival rate to the j^{th} server for $j \geq 1$

λ_0 is the external arrival rate

C_s^j is the squared coefficient of variation of the service time distribution in the j^{th} server

P_{ji} is the routing probability that, after departing from the j^{th} server, the job will join the i^{th} server queue, for $j \geq 1$

P_{0i} is the probability that the external arrival will join the i^{th} server queue

Notice $\{\lambda_i\}$ is the solution of the following system of linear equations

$$\lambda_i = P_{0i} + \sum_{j=1}^{n} \lambda_j P_{ji} \qquad (7.3)$$

A detailed interpretation of this equation can be found in [13]. For open queueing networks, (7.3) provides unique solution to $\{\lambda_i\}$. In this section, all the queueing networks considered are assumed to be open queueing networks. Extension to closed queueing networks follows the treatment in [12].

Gelenbe [6] argued that Reiser and Kobayashi [12] did not put into consideration of the effect of idle period on the variance of the interdeparture time and assumed that the interdeparture time of the i^{th} server, τ_i, is a service time S_i with probability ρ_i or an interarrival time, A_i, plus a service time with probability $(1-\rho_i)$.

By straightforward manipulation, we get

$$E\left\{\tau_i^2\right\} = E\left\{S_i^2\right\} + (1-\rho_i)\left(E\left\{A_i^2\right\} + 2E\left\{A_i\right\} E\left\{S_i\right\}\right) \qquad \text{for} \quad 1 \le i \le n$$

and by definition

$$E\left\{\tau_i^2\right\} = \lambda_i^{-2}(1 + C_i) \qquad \text{for} \quad 1 \le i \le n$$

Combining the two equations together, we get

$$C_i + 1 = \rho_i^2\left(C_s^i + 1\right) + (1-\rho_i)\left(\lambda_i^2 E\left\{A_i^2\right\} + 2\rho_i\right) \qquad \text{for} \quad 1 \le i \le n \qquad (7.4)$$

Finally, making the assumption that the number of arrivals forms a renewal process, we get

$$\sum_{j=0}^{n} \left[(C_j - 1)\, p_{ji} + 1 \right] \lambda_j p_{ji} = \lambda_i^3 \left(E\left\{A_i^2\right\} - \left(\lambda_i^{-1}\right)^2 \right) \qquad \text{for} \quad 1 \le i \le n$$

After simplification, we get the following system of linear equations:

$$C_i = \rho_i^2 (C_s^i + 1) + (1 - \rho_i) \left(2\rho_i + 1 + \lambda_i^{-1} \sum_{j=0}^{n} \left[(C_j - 1)\, p_{ji} + 1 \right] \lambda_j p_{ji} \right) - 1$$

$$\text{for} \quad 1 \le i \le n \tag{7.5}$$

The assumption that τ_i *is equal to service time,* S_i, with probability ρ_i or an interarrival time plus a service time, $A_i + S_i$, with probability $(1 - \rho_i)$ is exact only for Markovian queueing networks. In Markovian queueing networks, $E\{A_i^2\}$ is equal to $2/\lambda_i^2$. Hence, a simpler approach is to further approximate $E\{A_i^2\}$ by $2/\lambda_i^2$; then, we get from (7.4) that

$$C_i = \rho_i^2 \left(C_s^i - 1 \right) + 1 \tag{7.6}$$

That is to say, we can now directly express C_i in a closed form expression, and the necessity of solving a system of linear equations has been eliminated. The value obtained by both methods are very close, as we shall see.

Alternatively, we may consider the problem in the following way. Let $\mathcal{C}$ be the set of service centers whose customers after service completion may go to service center A for further service. Since the arrival rate to each service center in the open queueing network is known, we first unfold the network but retain the connections from those service centers in $\mathcal{C}$ to A. If A is in $\mathcal{C}$, i.e., there is self loop at A, a duplication of A is used to replace the self loop. Then, we apply a Poisson input to each service center in $\mathcal{C}$ with the same arrival rate as before and then calculate its coefficient of variation of interdeparture time. Finally, treating each departure process as an independent renewal process, we can obtain the coefficient of variations of the interarrival time at service center A in this case. The coefficient of variation obtained will be used as an approximation of that of the original network and is exactly the same as substituting (7.6) into (7.1). For example, in the network given in Fig. 7.1, the appropriate subnetwork for evaluating the coefficient of variation of the interarrival time of the service centers 1 and 2 are given in Figs. 7.2a and 7.2b, respectively.

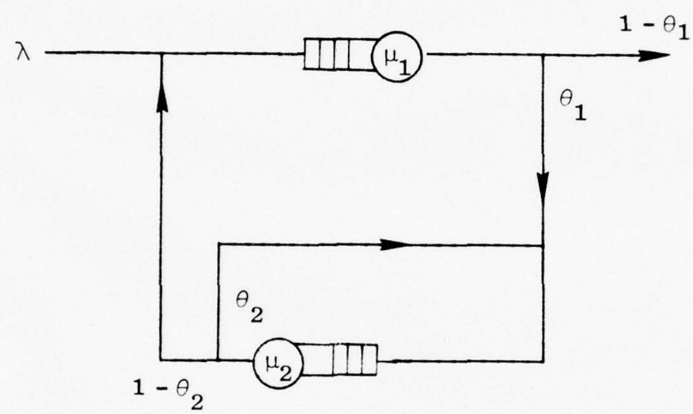
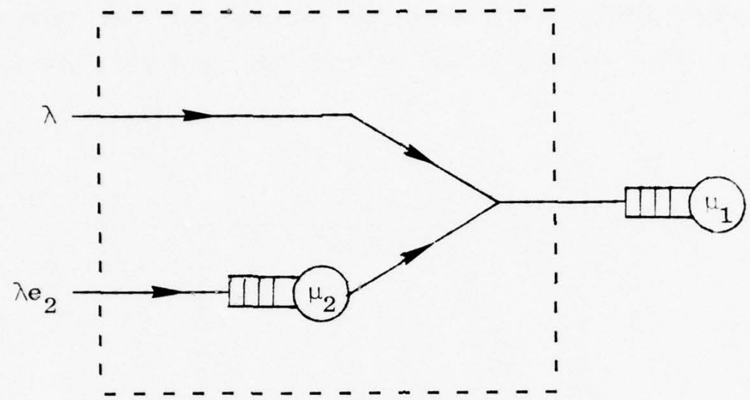
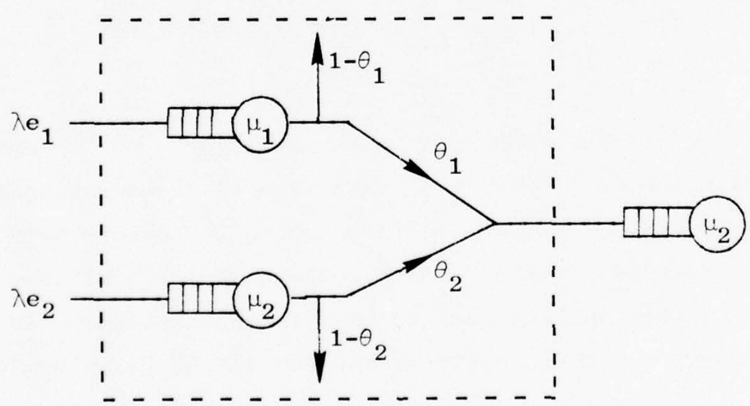


Fig. 7.1. OPEN TWO SERVER QUEUEING MODEL.

We now use the queueing network shown in Fig. 7.1 to compare the C_i's, the squared coefficients of variation of the interdeparture time at each server, by our straightforward method, Gelenbe's method, and Reiser and Kobayashi's method. Both authors [6,12] have used this network to demonstrate the accuracy of their approximations. In Table 7.1, the column under method P^* contains the results of the proposed method, and the column under method B^* contains the results of the method proposed by Gelenbe [6], and the column under method A^* contains the results of the method proposed by Reiser and Kobayashi [12]. As we can see, all three methods yield the same answer when the network is a Markovian queueing network. The estimates of the coefficient of variation of the interdeparture time obtained by our method and Gelenbe's method are always very close, as it should be since both methods try to incorporate the effect of idle period on the coefficient of variation of the interdeparture time. However, our method is much simpler in computation.



(a) First server



(b) Second server

Fig. 7.2. APPROXIMATE NETWORK CONFIGURATION FOR ESTIMATING THE COEFFICIENTS OF VARIATION OF INTERARRIVAL TIMES FOR THE NETWORK IN FIG. 7.1.

Table 7.1

THE SQUARED COEFFICIENTS OF VARIATIONS OF INTERDEPARTURE TIMES (C_1, C_2) OF THE QUEUEING NETWORK IN FIG. 7.1 WITH $\rho_1 = 0.9$, $\rho_2 = 0.84$

(a) $\theta_1 = \theta_2 = 0.5$

C_s^1	C_s^2	Method P^*		Method B^*		Method A^*	
		C_1	C_2	C_1	C_2	C_1	C_2
0.5	0	0.595	0.294	0.576	0.247	0.5	0
0.5	0.5	0.595	0.647	0.585	0.615	0.5	0.5
1.0	0.5	1.0	0.647	0.991	0.632	1.0	0.5
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

(b) $\theta_1 = 0.5$, $\theta_2 = 0$

C_s^1	C_s^2	Method P^*		Method B^*		Method A^*	
		C_1	C_2	C_1	C_2	C_1	C_2
1.0	0	1.0	0.294	0.965	0.292	1.0	0
1.0	0.25	1.0	0.471	0.973	0.469	1.0	0.25
1.0	0.5	1.0	0.647	0.982	0.646	1.0	0.5
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2.0	1.0	1.81	1.0	1.81	1.07	2.0	1.0
2.0	0.5	1.81	0.647	1.80	0.711	2.0	0.5

Let us apply the diffusion approximation techniques to analyze the computer communication network in Fig. 7.3. The network has the same topology as the communication network, CIGALE, within CYCLADES [28] which is a general purpose computer network being installed in France. All the terrestrial links are assumed to be full duplex. The numbers on the terrestrial links represent servers and their queues. Thus, 3 refers to the server which transfers messages from node C to node A and 2 refers to the server which transfers messages in the opposite direction. Traffic moving in the two opposite directions along the same link is assumed to be noninterfering. Each station receives external traffic which forms a Poisson process. We also assume that each message arriving from outside to station i has equal probabilities of having any of the other 4 stations as its final destination. The routing algorithm of the networks is assumed to be fixed and will be described later. All the above assumptions about the terrestrial network have been adopted by Gelenbe [6] in modeling the CYGALE network under packet switching.

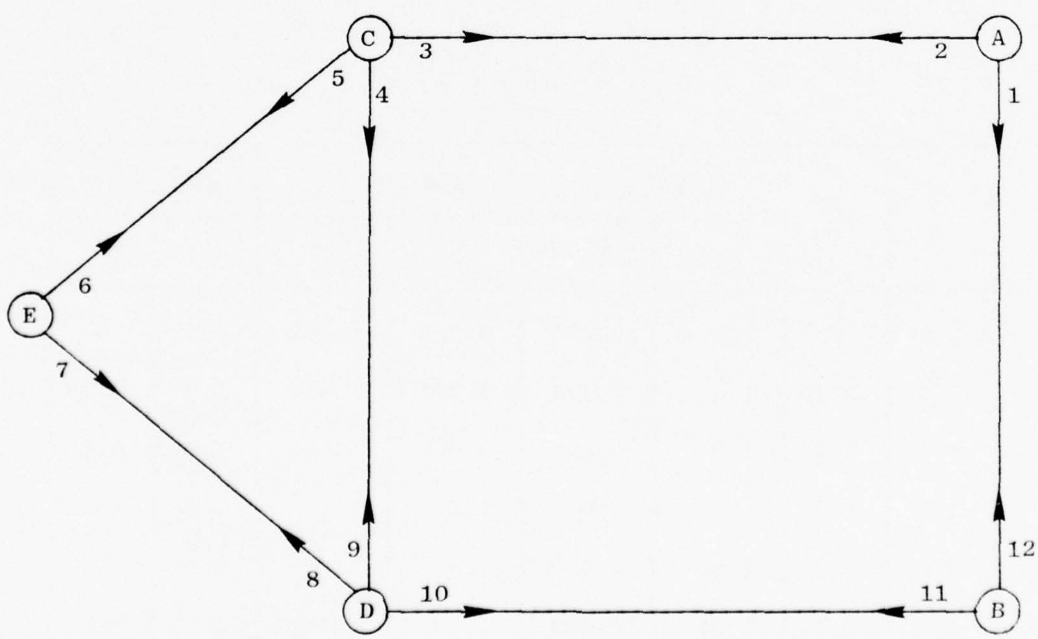


Fig. 7.3. COMPUTER COMMUNICATION NETWORK.

Let cH_i be the channel capacity, the number of packets that can be transmitted per second, of link i. The channel capacity of each link is indicated in Table 7.2. The fixed routing algorithm is summarized in Table 7.3. The routes which are not shown in Table 7.3 are the links which directly connect the source stations and destination stations. The number of packets contained in each message is assumed to be geometrically distributed with mean five. The external arrival rate at each node is tabulated in Table 7.4.

The performances of the network under both message switching and packet switching are analyzed. In Table 7.5a, mean queue lengths at each server under message switching by various approximation methods and simulation are tabulated. Again, our method is denoted by method P, Gelenbe's method is denoted by method B, and Reiser and Kobayashi's method is denoted by method A. The simulation results are presented with 95% confidence intervals. In Table 7.5b, the corresponding squared coefficients of variation of interdeparture time at each server by various methods are tabulated. Both methods P^* and B^* lead to similar results on the squared coefficients of variation of interdeparture times. The difference in mean queue by methods P and B in Table 7.5a is mainly due to different ways being employed in estimating diffusion parameters. The minor difference in estimating squared coefficient of variation of interdeparture time has very little effect. As we can see, method P leads to better approximation. In Table 7.6a, the mean queue lengths under packet switching obtained by methods P, B, A and simulation are tabulated. In Table 7.6b, the corresponding squared coefficient of variation of interdeparture time at each server by various methods are tabulated. Not only the squared coefficients of variation of interdeparture time but also the mean queue lengths obtained by method P and method B are very close to each other. Furthermore, the mean queue lengths obtained by both methods are very close to the simulation result. As we can see from Tables 7.5a and 7.6a, method P provides very accurate approximations in both cases, where other methods can provide very accurate approximations in only one of the two cases.

Finally, we consider the decomposability problem of general queueing networks. From our previous analyses in Sections 5 and 6, we expect that, if the service time distributions of some of the intermediate

Table 7.2

CHANNEL CAPACITY OF EACH LINK
IN THE TERRESTRIAL NETWORK

Link	cH_i (Packet/Sec)
1,12	50
2,3	80
4,9	70
5,6	45
7,8	50
10,11	70

Table 7.3

ROUTING TABLE

Source Stations	Destination Stations	Route
A	D	2,4
A	E	2,5
B	C	12,2
B	E	11,8
C	B	4,10
D	A	9,3
E	A	6,3
E	B	7,10

Table 7.4

EXTERNAL ARRIVAL RATE (PER SECOND)

Node	Message Arrival Rate	Packet Arrival Rate
A	12	60
B	16	80
C	16	80
D	16	80
E	16	80

Table 7.5a

MEAN QUEUE LENGTH UNDER MESSAGE SWITCHING

Server	Simulation	Method P	Method B	Method A
1	0.416*	0.416	0.536	0.417
2	3.895 ± 0.080	3.947	4.272	3.862
3	2.732 ± 0.092	2.733	3.033	2.646
4	3.356 ± 0.081	3.367	3.681	3.300
5	3.225 ± 0.076	3.210	3.521	3.141
6	7.289*	7.289	7.644	7.210
7	3.680*	3.680	4.000	3.617
8	3.606 ± 0.165	3.654	3.974	3.537
9	1.257*	1.257	1.486	1.230
10	5.488 ± 0.225	5.392	5.735	5.265
11	1.257*	1.257	1.486	1.230
12	3.680*	3.680	4.000	3.617

*Exact.

Table 7.5b

SQUARED COEFFICIENTS OF VARIATION OF INTERDE-
PARTURE TIME UNDER MESSAGE SWITCHING

Server	Method P*	Method B*	Method A*
1	0.982	0.982	0.8
2	0.868	0.864	0.8
3	0.888	0.878	0.8
4	0.877	0.875	0.8
5	0.879	0.876	0.8
6	0.842	0.842	0.8
7	0.872	0.872	0.8
8	0.872	0.869	0.8
9	0.935	0.935	0.8
10	0.853	0.848	0.8
11	0.935	0.935	0.8
12	0.872	0.872	0.8

Table 7.6a

MEAN QUEUE LENGTH UNDER PACKET SWITCHING

Server	Simulation	Method P	Method B	Method A
1	0.364*	0.364	0.364	0.303
2	2.399 ± 0.094	2.400	2.400	1.933
3	1.621 ± 0.051	1.666	1.666	1.186
4	2.165 ± 0.061	2.166	2.165	1.781
5	2.027 ± 0.063	2.050	2.048	1.656
6	4.444*	4.444	4.444	4.018
7	2.400*	2.400	2.400	2.033
8	2.301 ± 0.112	2.269	2.269	1.644
9	0.952*	9.952	0.952	0.736
10	2.881 ± 0.116	2.962	2.959	2.293
11	0.952*	0.952	0.952	0.736
12	2.400*	2.400	2.400	2.033

*Exact.

Table 7.6b

SQUARED COEFFICIENTS OF VARIATION OF INTERDEPARTURE TIME UNDER PACKET SWITCHING

Server	Method P*	Method B*	Method A*
1	0.91	0.91	0
2	0.340	0.321	0
3	0.438	0.391	0
4	0.383	0.374	0
5	0.395	0.380	0
6	0.210	0.210	0
7	0.360	0.360	0
8	0.360	0.343	0
9	0.673	0.673	0
10	0.265	0.239	0
11	0.673	0.673	0
12	0.360	0.360	0

servers or the external interarrival time distributions not only have large coefficients of variation but also are type B hyperexponential distributions, the diffusion approximation will not work. This is simply because the arrival processes of the servers receiving jobs from those servers with high coefficients of variation of service times will have high coefficients of variation. However, even if the service time distributions with high coefficients of variation are type A hyperexponential distributions, the diffusion approximations under decomposition techniques still may not be satisfactory for certain network topology. The most noteworthy networks of this type are network with feedback loops, especially self loops. Let us take a second look at the network in Fig. 7.1. When the coefficients of variation of the service times in both servers are not large, the decomposability of the network seems to be acceptable from the results obtained by various diffusion approximations in [6] and [12].

Now, let the service time distribution of the first server be hyperexponential distribution with the following parameter:

$$\omega = 0.029249$$

$$M_1 = 30.000187$$

$$M_2 = 0.0232795$$

Recall the hyperexponential density function has the form $(\omega/M_1)\, e^{-X/M_1} + ((1-\omega)/M_2)\, e^{-X/M_2}$. This hyperexponential distribution has mean 0.9 and sq. coeff. of variation 64. One of its branches has mean very close to zero. Let the second server have exponential service time with mean 0.84. Furthermore, let

$$\theta_1 = 0.5$$

$$\theta_2 = 0.5$$

$$\lambda_0 = 0.5$$

The simulation result of the mean queue length in the first server with 95% confidence interval is 165 ± 15. The mean queue lengths obtained by diffusion approximation are around 294 under methods A and B and 264 under method P. That is to say, all methods tend to overestimate the mean queue length. We now try to give a reasonable explanation to this anomaly. Recall the arrival rate to the first server can be calculated by solving the system of linear equations (7.3). After simple manipulation, we get $\lambda_1 = 1$. Notice the external arrival rate is 0.5, so one-half of the arrivals to the first server is from the feedback path through the second server. When the first server encounters a long service time, the arrival process from the second server will be shut down after the second queue becomes empty. That is to say, the arrival rate is effectively 0.5 instead of 1 during the later period of the long service time. However, in the diffusion approximations, the arrival rate to the first server is always 1. This is the reason why the mean queue length is overestimated under diffusion approximation. The correlations of the service stations become very serious as the coefficients of variation of service time distributions become large in this type of network, and using the ordinary way to estimate the diffusion parameters is not sufficient to account for this sort of correlations. What will happen if the service time distribution is a type B hyperexponential distribution function indicated in Section 5? A simulation has been conducted when the first server has a two stage hyperexponential distribution with the following parameters:

$$\omega = 0.000126$$

$$M_1 = 500$$

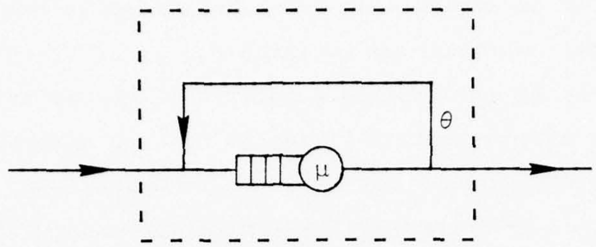
$$M_2 = 0.843486$$

The second server still has exponential distribution. The network topology and the traffic intensity is the same as the previous example. Hence, the mean queue length predicted by diffusion approximation is still the same. But the 95% interval estimation obtained by simulation is 96 ± 27, which is quite small, as expected. The broad width of the confidence interval is due to the heavy traffic condition and the closeness of ω

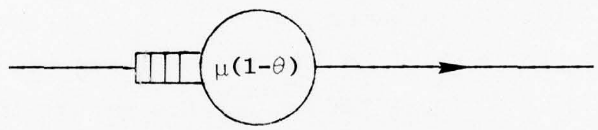
to 0. The number of arrivals in the simulations is around 10^6, hence the point estimation should be acceptable.

So, the idea of decomposing a network of queues into separate single server queueing systems does not seem to be always feasible. In certain network topologies, such as network with feedback loops, especially self loops, there is a dependence of the arrival process of each service center in the feedback path on its departure process. If the service center has a self loop which contributes a large portion of arrivals, the effect of this dependency becomes very serious as the coefficient of variation of the service time deviates significantly from 1. Using any of the three methods cited above to estimate the coefficient of variation of the interarrival time, the decomposition technique can not reflect this dependency into the estimated parameter.

To be more specific, let us consider method P. This fact can be observed from Fig. 7.2b where the two service centers with rate μ_2 are actually the same but are represented as two different service centers. Clearly, the dependence among the two is not reflected in the estimated parameters. Nevertheless, the self loop problem can be solved by treating the server and its self loop as a single entity as we did in Section 6. That is to say, we first eliminate all self loops in the network by replacing each server with a self loop by an equivalent server without a self loop as in Fig. 7.4 and then apply the decomposition technique if possible. Let β' be the contribution to the diffusion parameter β from the server and its self loop if any. In Table 7.7, we tabulate the approximate values of β' under methods P* and A* when the self loop is not eliminated and the correct value under the equivalent server without a self loop. The error terms under the two approximation methods are proportional to $C_s(\theta^2+\theta)$. This explains why direct decomposition does not work for a strong self loop under a large coefficient of variation of service time even if the distribution is type A. Besides the anomalies due to type B distributions, the problem still not solved is strong feedback loops which are not self loops under large coefficients of variation of service times. Although the analysis is greatly simplified when decomposition does work, decomposition is not a panacea. We should be careful about the decomposability of the queueing network and all the distributions involved.



(a) Server with a self loop
Variance of service time: σ_s^2



(b) Equivalent server without a self loop
Variance of service time: $\sigma_s^2/(1-\theta) + \theta/(\mu(1-\theta))^2$

Fig. 7.4. SERVER WITH A SELF LOOP AND ITS EQUIVALENT REPRESENTATION WITHOUT A SELF LOOP.

Table 7.7

β' UNDER VARIOUS METHODS

β'	Without a Self Loop	With a Self Loop	
		Method A*	Method P*
	$\lambda(C_s + \theta - \theta C_s)$	$\lambda(C_s + \theta + \theta^2(C_s - 1))$	$\lambda(C_s + \theta + \theta^2 \rho^2(C_s - 1))$

8. THE SERVICE CENTER WITH A QUEUE DEPENDENT SERVICE RATE OR ARRIVAL RATE

In this section, we consider the case where a service center has queue dependent service rate. The conventional G/M/m queueing system is a special case of this class of service centers. The service rate of the m-server queueing system can be expressed as

$$\mu_i = \begin{cases} i\mu & \text{for } i < m \\ m\mu & \text{for } i \geq m \end{cases}$$

where i is the number of customers in the queue.

In computer system modeling, a more general μ_i than the conventional one cited above is often needed. Consider the performance of a tightly coupled computer system. The total service rate of CPU's does not increase as a linear function of the number of CPU's in the system due to the memory interference among different CPU's.

When the service rate is queue dependent, the diffusion parameters also become queue dependent and the diffusion equation becomes harder to solve. We first need to determine the values of the diffusion parameters at those integer points and then propose a reasonable way to interpolate their values in between integers. The infinite capacity case is first considered. We further assume that μ_i will keep constant for $i \geq m$, as in the conventional multiserver case with m servers. Similarly, for the arrival rate λ_i.

The values of the diffusion parameters at those integer points are defined as

$$\begin{cases} \beta_i = \lambda_i - \mu_i \\ \alpha_i = C_a\lambda_i + C_s g\mu_i \end{cases} \tag{8.1}$$

and r_i, as usual, is defined to be

$$r_i = \frac{2\beta_i}{\alpha_i}$$

where C_a and C_s are the squared coefficients of variation of the interarrival time and service time distributions, respectively. When the service rate is fixed, we set g equal to the traffic intensity of the queueing system. But, for a server with queue dependent service rate or arrival rate, the appropriate value of g is not very clear. From experimental results, it seems to be that, if we set g equal to 1 for the case where $C_a \leq 1/2$ and $C_s \leq 1$ and λ_m/μ_m otherwise, the approximation will be more accurate in general.

There are at least two different ways to interpolate the value of $\alpha(X)$ and $\beta(X)$ in between integers.

Method 1. Interpolation by Step Functions (see Fig. 8.1)

$$\alpha(X) = \begin{cases} \alpha_1 & X \leq \frac{3}{2} \\ \alpha_k & k - \frac{1}{2} < X \leq k + \frac{1}{2} \text{ and } 2 \leq k \leq m - 1 \\ \alpha_m & X > m - \frac{1}{2} \end{cases} \tag{8.2}$$

Similarly for $\beta(X)$.

Method 2. Interpolation by Linear Functions (see Fig. 8.2)

$$\alpha(X) = \begin{cases} \alpha_1 X & X \leq 1 \\ \alpha_k + (\alpha_{k+1} - \alpha_k)(X - k) & k < X \leq k + 1 \text{ and } 1 \leq k \leq m - 1 \\ \alpha_m & X > m \end{cases} \tag{8.3}$$

or let $\alpha(X) = \alpha_1$ for $X \leq 1$ to make the impulse term $\delta(X - 1)$ in the diffusion equation easier to handle. Similarly for $\beta(X)$.

Again, we use Feller's elementary return process to handle boundary condition. The diffusion equation satisfied by the probability density

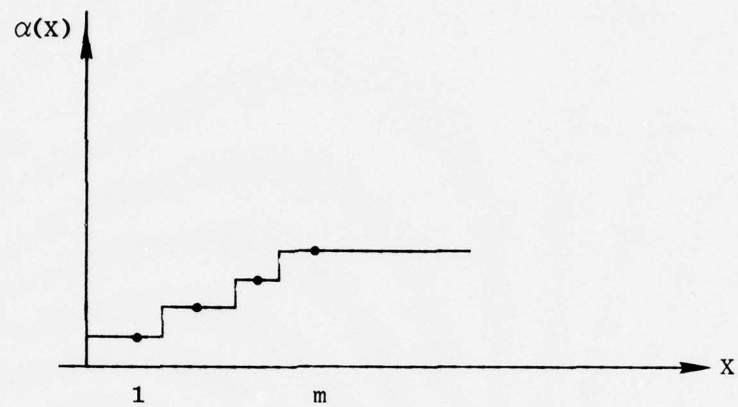


Fig. 8.1. INTERPOLATION OF $\alpha(X)$ USING STEP FUNCTIONS.

Fig. 8.2. INTERPOLATION OF $\alpha(X)$ USING LINEAR FUNCTIONS.

function $P(X)$ of the approximate queue length process has the following form under steady state

$$\frac{1}{2}\frac{d^2}{dX^2}\alpha(X)\,P(X) - \frac{d}{dX}\beta(X)\,P(X) = -\lambda_0 M_1 \delta(X-1) \tag{8.4}$$

with boundary conditions

$$\lim_{X\to 0} \frac{1}{2}\frac{d}{dX}\alpha(X)\,P(X) - \beta(X)\,P(X) = \lambda_0 M_1$$

and

$$P(0) = 0$$

where M_1 is the probability that the queueing system is idle.

Under the second interpolation method, numerical integration is required to estimate the queue length distribution. For the conventional multiserver, the broken line in Fig. 8.2 becomes a straight line and the complexity of the problem is simplified, but numerical integration is still needed. Hence, for better mathematical tractability, we adopt the first interpolation method. Halachmi and Franta [26,25] have applied diffusion approximation to conventional multiserver queueing systems using the second interpolation method and reflection boundary for both infinite and finite population models, respectively. The results from both methods seem to be quite close in the few cases examined.

After solving the differential equation, we get

$$P(X) = \begin{cases} \dfrac{\lambda_0 M_1}{\beta_1}\left(e^{Xr_1} - 1\right) & \text{for } 0 \le X \le 1 \\[2ex] M_1 d\, e^{(X-1)r_1} & \text{for } 1 < X \le \frac{3}{2} \\[2ex] M_1 d\, e^{S_{k-1} + (X-k+\frac{1}{2})r_k} & \text{for } k-\frac{1}{2} < X \le k+\frac{1}{2} \text{ and } 2 \le k \le m-1 \\[2ex] M_1 d\, e^{S_{m-1} + (X-m+\frac{1}{2})r_m} & \text{for } X > m-\frac{1}{2} \end{cases} \tag{8.5}$$

where

$$S_k = \sum_{i=1}^{k} r_k - \frac{r_1}{2} \qquad \text{for} \quad 1 \leq k \leq m - 1 \tag{8.6}$$

and

$$d = \frac{\lambda_0}{\beta_1}\left(e^{r_1} - 1\right)$$

The unknown constant M_1 can be determined by the fact that total probability must sum up to 1, i.e.,

$$\int_0^{\infty} P(X)\, dX + M_1 = 1$$

After simplification, we get

$$M_1 = \left(1 + \frac{\lambda_0}{\beta_1 r_1}\left(e^{r_1} - 1 - r_1\right) + \frac{d}{r_1}\left(e^{r_1/2} - 1\right)\right.$$

$$\left. + \sum_{k=1}^{m-1} \frac{d}{r_k}\, e^{S_{k-1}}\left(e^{r_k} - 1\right) - \frac{d}{r_m}\, e^{S_{m-1}}\right)^{-1} \tag{8.7}$$

If r_k is equal to zero, we should replace the term $(e^{r_k} - 1)/r_k$ by 1. Similar remark holds for the rest of the section.

Let π_i be the probability that i jobs are in the system. We define

$$\pi_0 = M_0$$

$$\pi_1 = \int_0^{3/2} P(X)\, dX$$

$$\pi_k = \int_{k-1/2}^{k+1/2} P(X)\, dX \qquad \text{for} \quad k \geq 2$$

$$\pi_k = \begin{cases} \dfrac{\lambda_0 M_1}{\beta_1 r_1}\left(e^{r_1} - 1 - r_1\right) & k = 1 \\ \dfrac{M_1 d}{r_k} e^{S_{k-1}}\left(e^{r_k} - 1\right) & 2 \le k \le m - 1 \\ \dfrac{M_1 d}{r_m} e^{S_{m-1}+(k-m)r_m}\left(e^{r_m} - 1\right) & k \ge m \end{cases} \tag{8.8}$$

In Table 8.1, we compare the mean queue lengths obtained under diffusion approximation with analytic results for the M/M/m system with $m = 2, 3, 4, 5, 6, 7,$ and 8, when $\rho \triangleq \lambda/m\mu = 0.95$ and 0.85, respectively. The approximation is very accurate. Then, we compare the conditional mean queue length of the external queue (given that external queue exists, i.e., number of jobs in the system is larger than m) with the analytic result for the G/M/m queueing system. Both the analytic result and diffusion approximation on the conditional mean external queue length of the G/M/m queueing system are independent of m. After simple manipulation, we can get the conditional mean external queue length under diffusion approximation which is $1/(1 - e^{r_m})$. The exact result is $1/(1 - \sigma)$ where σ is defined in Section 3. In Table 8.2, we compare the conditional mean external queue length obtained under diffusion approximation with the analytic result for $E_2/M/m$ and $E_3/M/m$ queueing systems when $\rho = 0.95, 0.90, 0.85, 0.80,$ and 0.75.

When the arrival process has hyperexponential distribution, again, the conditional mean external queue length can vary over a wide range. Nevertheless, for type A hyperexponential distribution, diffusion approximation can still be applied as before. Tables 8.3 and 8.4 tabulate the diffusion approximations and analytic results under different values of M_2 for $C_a = 2, 4, 8, 16,$ and 32, when $\rho = 0.95$ and 0.85, respectively. In both cases, λ is equal to 1.

In Table 8.6, we consider the case where μ_i is an arbitrary function of i and λ_i is constant. The result is again satisfactory.

We now consider the closed two server queueing network in Fig. 8.3a which can be interpreted as the CPU/DTU model, as noted earlier.

Table 8.1

MEAN QUEUE LENGTHS FOR M/M/m SYSTEM WHEN $\rho = 0.85$ AND 0.95

m	ρ = 0.85		ρ = 0.95	
	Diffusion Approximation	Exact	Diffusion Approximation	Exact
2	6.031	6.126	19.37	19.49
3	6.695	6.689	20.07	20.08
4	7.354	7.306	20.75	20.74
5	8.028	7.959	21.46	21.43
6	8.718	8.636	22.18	22.15
7	9.425	9.333	22.92	22.88
8	10.14	10.04	23.68	23.64

Table 8.2

CONDITIONAL MEAN EXTERNAL QUEUE LENGTH
FOR E_2/M/m AND E_3/M/m SYSTEM

ρ	E_3/M/m		E_2/M/m	
	Diffusion	Exact	Diffusion	Exact
0.95	13.67	13.45	15.26	15.09
0.90	7.013	6.784	7.761	7.588
0.85	4.797	4.566	5.268	5.091
0.80	3.693	3.460	4.023	3.844
0.75	3.033	2.797	3.280	3.097

Table 8.3

CONDITIONAL MEAN EXTERNAL QUEUE LENGTH FOR $H_2/M/m$ SYSTEM WHEN $\rho = 0.95$

C_a	Diffusion	Exact	
		$M_2 = 0.2$	$M_2 = 0.7$
2	29	29.9	29.4
4	48	49.7	47.5
8	86	89.6	83.0
16	162	169	153
32	314	316	294

Table 8.4

CONDITIONAL MEAN EXTERNAL QUEUE LENGTH FOR $H_2/M/m$ SYSTEM WHEN $\rho = 0.85$

C_a	Diffusion	Exact	
		$M_2 = 0.1$	$M_2 = 0.6$
2	9.01	9.94	9.53
4	14.7	16.5	14.8
8	26.0	29.5	24.9
16	48.7	55.8	44.8
32	94.0	108	84.1

AD-A042 722 STANFORD UNIV CALIF STANFORD ELECTRONICS LABS F/G 9/2
ON ACCURACY IMPROVEMENT AND APPLICABILITY CONDITIONS OF DIFFUSI--ETC(U)
JAN 77 P S YU DASG60-77-C-0073
UNCLASSIFIED SU-SEL-77-016 NL

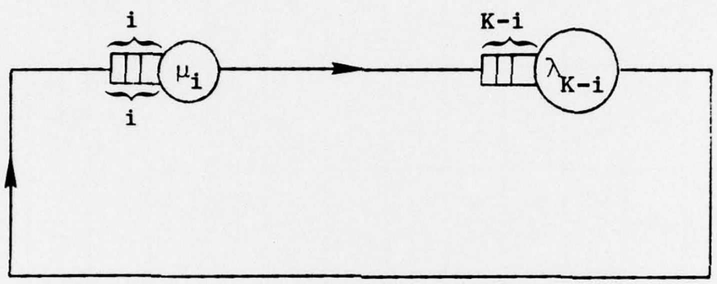
2 OF 2
ADA042 722

END
DATE
FILMED
9-77
DDC

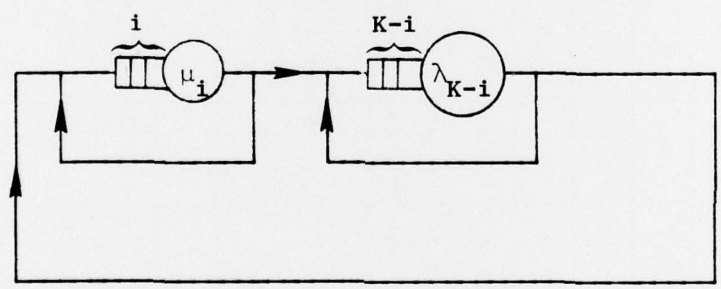
Table 8.5

MEAN QUEUE LENGTH FOR SERVER WITH GENERAL QUEUE DEPENDENT SERVICE RATE

m	3	3	3	3	5	5	5	5	6	6	7	7
λ	2.55	2.85	2.55	2.85	4.25	4.75	4.25	4.75	5.95	6.65	5.95	6.65
μ_1	3.0	3.1	1.0	1.0	5.0	5.5	1.0	1.0	6.5	7.2	1.0	1.0
μ_2	2.9	2.9	1.8	1.8	4.8	5.4	1.8	1.8	6.4	7.1	1.9	1.9
μ_3	2.6	2.9	2.6	3.0	4.6	5.2	2.6	2.6	6.3	7.0	2.8	2.8
μ_4					4.5	5.0	3.5	3.5	6.2	6.9	3.7	3.7
μ_5					4.4	4.8	4.4	5.0	6.1	6.8	4.6	4.6
μ_6									6.0	6.7	5.5	5.5
μ_7											6.4	7.0
Mean Queue Length												
Diffusion	49.91	56.38	51.99	20.16	27.11	93.34	30.81	21.75	117.7	131.8	17.09	23.23
Exact	50.61	56.93	52.03	20.17	27.80	94.14	30.79	21.72	118.4	132.5	17.04	23.19



(a) Model without a self loop



(b) Model with a self loop at each server

Fig. 8.3. CPU-DTU MODEL WITH K DEGREE OF MULTIPROGRAMMING.

Generalization to the model in Fig. 8.3b follows the same idea cited in Section 6. That is to say, we first compute the mean and variance of the service time of the equivalent server without a self loop. The only difference from before is that the quantities now are queue dependent. After replacing each server and its self loop by an equivalent server without a self loop, the model in Fig. 8.3b reduces to that in Fig. 8.3a. The number of jobs in the system is assumed to be K. The queue dependent service rate of CPU and DTU are denoted by μ_i and λ_{K-i} when there are i jobs in the CPU queue. The diffusion parameters are defined to be

$$\begin{cases} \beta_i = \lambda_{K-i} - \mu_i \\ \alpha_i = C_a \lambda_{K-i} + C_s \mu_i \qquad \text{for} \quad 1 \leq i \leq K \end{cases} \tag{8.9}$$

and, as before,

$$r_i = 2\beta_i / \alpha_i$$

Notice the definition is similar to the previous one with $g = 1$. Again, there are at least two different ways to interpolate the value of $\alpha(X)$ and $\beta(X)$ in between integers. The only difference is that now we have two boundaries. We still adopt the interpolation method using step functions for simplicity. To be more precise, we define

$$\alpha(X) = \begin{cases} \alpha_1 & 0 < X < \frac{3}{2} \\ \alpha_i & i - \frac{1}{2} < X < i + \frac{1}{2} \qquad 2 \leq i \leq K-2 \\ \alpha_{K-1} & K - \frac{3}{2} < X < K \end{cases} \tag{8.10}$$

Similarly for $\beta(X)$.

The diffusion equation now has the following form.

$$\frac{1}{2}\frac{d^2}{dX^2}\alpha(X)\ P(X) - \frac{d}{dX}\beta(X)\ P(X) = -\lambda_0 M_1 \delta(X-1) - \mu_K M_2 \delta(X - M + 1)$$

$$\lim_{X\to 0} \frac{1}{2}\frac{d}{dX}\alpha(X)\ P(X) - \beta(X)\ P(X) = \lambda_0 M_1 \tag{8.11}$$

$$\lim_{X\to K} \frac{1}{2}\frac{d}{dX}\alpha(X)\ P(X) - \beta(X)\ P(X) = \mu_K M_2$$

where M_2 is the probability that the CPU queue has K jobs.

After solving the diffusion equation (8.11), we get

$$P(X) = \begin{cases} \dfrac{\lambda_0 M_1}{\beta_1}\left(e^{Xr_1} - 1\right) & \text{for } 0 \le X \le 1 \\[2ex] M_1 d\ e^{(X-1)r_1} & \text{for } 1 < X \le \frac{3}{2} \\[2ex] M_1 d\ e^{S_{i-1}+(X-i+\frac{1}{2})r_i} & \text{for } i - \frac{1}{2} < X \le \max\ (K-1,\ i+\frac{1}{2}) \\ & \text{and } 2 < i \le m - 1 \\[2ex] M_1 d\ e^{S_{K-2}+\frac{r_K}{2}}\left(\dfrac{e^{r_{K-1}} - e^{r_{K-1}(X-K+1)}}{e^{r_{K-1}} - 1}\right) & \text{for } X > K - 1 \end{cases} \tag{8.12}$$

where

$$M_2 = \frac{M_1 d\beta_{K-1}}{\mu_K\left(e^{r_{K-1}} - 1\right)}\ e^{S_{K-2}+\frac{3}{2}r_{K-1}} \tag{8.13}$$

and S_i is defined as before in (8.6).

The unknown constant M_1 can be determined by the fact that total probability must sum up to one, i.e.,

$$\int_0^{\infty} P(X)\, dX + M_1 + M_2 = 1$$

After simplification, we get

$$M_1 = \left(1 + \frac{\lambda_0}{\beta_1 r_1} \left(e^{r_1} - 1 - r_1 \right) + \frac{d}{r_1} \left(e^{r_1/2} - 1 \right) \right.$$

$$+ \sum_{i=2}^{K-2} \frac{d}{r_i} e^{S_{i-1}} \left(e^{r_i} - 1 \right) + \frac{d}{r_{K-1}} e^{S_{K-2}} \left(e^{r_{K-1}/2} - 1 \right) \tag{8.14}$$

$$\left. + d\, e^{S_{K-2} + \frac{1}{2} r_{K-1}} \left(\frac{(r_{K-1} - 1)\, e^{r_{K-1}} + 1}{r_{K-1} \left(e^{r_{K-1}} - 1 \right)} \right) + \frac{d\beta_{K-1}}{\mu_K \left(e^{r_{K-1}} - 1 \right)} e^{S_{K-2} + \frac{3}{2} r_{K-1}} \right)^{-1}$$

Let π_i be the probability that i jobs are in the CPU. We define

$$\begin{cases} \pi_0 = M_1 \\ \pi_1 = \int_0^{3/2} P(X)\, dX \\ \pi_i = \int_{i-1/2}^{i+1/2} P(X)\, dX \\ \pi_{K-1} = \int_{K-3/2}^{K} P(X)\, dX \\ \pi_k = M_2 \end{cases}$$

In Table 8.6, we compare the mean queue length at the CPU when both CPU and DTU have exponential service time distributions, and furthermore the CPU is modeled as a traditional m-server under fixed degree of multiprogramming K. In Table 8.7, the case where service rate of CPU is an

Table 8.6

MEAN QUEUE LENGTH WHEN CPU IS
MODELED AS CONVENTIONAL m SERVER

m	K	$\rho = 0.95$		$\rho = 0.85$	
		Diffusion	Exact	Diffusion	Exact
2	4	2.28	2.13	2.11	1.94
3	3	2.00	1.92	1.91	1.81
3	5	3.09	2.89	2.88	2.64
4	4	2.81	2.69	2.70	2.55
4	7	4.38	4.14	4.02	3.73
5	5	3.64	3.49	3.48	3.31
5	8	5.19	4.94	4.78	4.48
6	6	4.47	4.31	4.27	4.08
6	10	6.48	6.22	5.89	5.56
7	7	5.30	5.14	5.07	4.87
7	11	7.31	7.05	6.66	6.33
8	8	6.15	5.98	5.87	5.66
8	12	8.15	7.89	7.44	7.11

$\rho = \lambda/m\mu$

Table 8.7

MEAN QUEUE LENGTH AND UTILIZATION AT CPU WHEN CPU HAS GENERAL QUEUE DEPENDENT SERVICE RATE

K	3	3	3	3	5	5	5	5	7	7	7	7
λ	2.55	2.85	2.55	2.85	4.25	4.75	4.25	4.75	5.95	6.65	5.95	6.65
CPU Service Rate												
μ_1	3.0	3.1	1.0	1.0	5.0	5.5	1.0	1.0	6.5	7.2	1.0	1.0
μ_2	2.9	2.9	1.8	1.8	4.8	5.4	1.8	1.8	6.4	7.1	1.9	1.9
μ_3	2.6	2.9	2.6	3.0	4.6	5.2	2.6	2.6	6.3	7.0	2.8	2.8
μ_4					4.5	5.0	3.5	3.5	6.2	6.9	3.7	3.7
μ_5					4.4	4.8	4.4	5.0	6.1	6.8	4.6	4.6
μ_6									6.0	6.7	5.5	5.5
μ_7									5.9	6.7	6.4	7.0
Mean Queue Length												
Diffusion	1.357	1.436	1.979	2.016	2.195	2.207	3.653	3.745	3.248	3.268	5.241	5.420
Exact	1.364	1.453	1.906	1.955	2.231	2.239	3.520	3.639	3.279	3.296	5.071	5.288
Utilization												
Diffusion	0.6971	0.7261	0.9150	0.9257	0.7763	0.7794	0.9871	0.9902	0.8455	0.8485	0.9973	0.9982
Exact	0.6997	0.7305	0.9066	0.9209	0.7821	0.7840	0.9859	0.9902	0.8483	0.8509	0.9974	0.9985

arbitrary function of queue length and that of DTU is constant is considered. Both the mean queue length and utilization at the CPU is tabulated. Again, the result is quite satisfactory.

9. CONCLUSION

Diffusion approximation is an attractive means of approximating the performance of queueing systems. In this paper, we not only assess the accuracy of diffusion approximation but also its limitation and applicable range. Modern computer systems are so complicated that oversimplified models may not predict any useful results. Realistic models often are not analytically tractable. Finding approximate solutions or upper bounds and lower bounds of the solutions is the only means to handle more complicated problems short of simulation. Under heavy traffic conditions, simulation converges very slowly and diffusion approximation seems to be the most attractive way to solve the problem. Nevertheless, diffusion approximation is not a panacea, it does have a limitation. This limitation has been overlooked in the past. Substantial effort has been devoted in this paper to identify the conditions where diffusion approximation can obtain accurate estimates. We must be careful with these conditions when applying diffusion approximation.

In Table 9.1, we classify the single server queueing systems according to their coefficients of variation of service times and interarrival times, and point out the diffusion approximation technique which seems to be most accurate according to our analysis. The superiority of method P, the proposed method, should be very apparent. When C_a is larger than one, the mean queue length may vary over a wide range even if the first two moments of interarrival time are kept constant. Diffusion approximation is applicable under the condition that the high variation of interarrival time is due to a great number of short interarrival times instead of a few very long interarrival times. Case studies have been conducted on 2-stage hyperexponential distributions which are widely used in computer system modelling. A similar anomaly is observed in two server closed queueing networks, often referred to as CPU/DTU models, when the service time of any server has a large coefficient of variation. Again, a similar regularity condition on service time distributions is required in order for the diffusion approximation to be applicable. Although method B does not yield the best performance when applying it to approximate the single server system, it is indeed a nice way to approximate

Table 9.1

RECOMMENDED DIFFUSION APPROXIMATION METHOD FOR SINGLE SERVER SYSTEM

	C_s Close to or Less than 0.5	C_s Close to 1	$C_s > 1$
$C_a < 1$	Method A	Method P	Method P
$C_a \approx 1$	Method P	Method P	Method P
$C_a > 1$*	Any Method		Method P (?)

*The high coefficient of variation of interarrival time must be due to a large number of short interarrival times instead of a few very long interarrival times.

the two server closed queueing network. When the coefficient of variation of the CPU service time is small, method G1 has similar performance. As the coefficient of variation of the CPU service time increases, method P becomes somewhat better.

For general queueing networks, an efficient way of taking into account the effect of idle periods to estimate the coefficient of variation of the arrival process at each server when the network can be decomposed into separate single servers is proposed. For certain network topologies, the arrival processes of some service centers strongly depend upon their own departure processes. Networks of this type are networks with strong feedback loops, especially self loops. When the coefficients of variation of the service times at the service centers have a large deviation from one, this sort of queueing network can not be decomposed into separate single servers directly. This fact has been neglected in the past. Nevertheless, the self loop problem can be solved by replacing each server with a self loop by an equivalent server without a self loop. After eliminating all the self loops, we can reconsider the decomposition of a network. The problem still not solved seems to be networks with strong feedback loops which are not self loops when the coefficients of variation of some service times are large. Surely, the regularity condition that a large coefficient of variation of external interarrival time or service time of each intermediate server is due to a lot of short interarrival times or service times, respectively, must always hold in order for diffusion approximation to be applicable.

Finally, we consider the service center with queue dependent service rate or arrival rate. General queue dependent service rate is often encountered in computer system modeling. Generalization to closed two server queueing network where each server may have a self loop is also considered.

REFERENCES

1. W. Feller, "Diffusion Processes in One Dimension," Trans. Am. Math. Soc. 77, 1954, pp. 1-31.

2. D. P. Gaver and G. S. Shedler, "Processor Utilization in Multiprogramming Systems Via Diffusion Approximations," Operation Research 21, 1973.

3. D. P. Gaver and G. S. Shedler, "Approximate Models for Processor Utilization in Multiprogrammed Computer Systems," SIAM J. Computing 2, Sept. 1973, pp. 183-192.

4. D. P. Gaver, Jr., "Diffusion Approximations and Models for Certain Congestion Problems," J.A.P. 5, 1968.

5. E. Gelenbe, "On Approximate Computer System Models," JACM 22, April 1975, pp. 261-269.

6. E. Gelenbe and G. Pujolle, "The Behavior of a Single Queue in a General Queueing Network," Acta Informatica 7, 1976.

7. E. Gelenbe, "A Non-Markovian Diffusion Model and Its application to the Approximation of Queueing System Behavior," T.R. 158, RIA, March 1976.

8. J. F. C. Kingman, "The Heavy Traffic Approximation in the Theory of Queues," in Proceedings of the Symposium on Congestion Theory, W. L. Smith and W. E. Wilkinson (ed), U. of North Carolina Press, 1965, pp. 137-169.

9. L. Kleinrock, "Queueing Systems," Volume I, II, Wiley Interscience (1975).

10. H. Kobayashi, "Application of the Diffusion Approximation to Queueing Networks, I, II," JACM 21, 1974, pp. 316-328, 459-469.

11. G.F. Newell, "Applications of Queueing Theory," 1971, Chapman and Hall, Ltd., (1971).

12. M. Reiser and H. Kobayashi, "Accuracy of the Diffusion Approximation for Some Queueing Systems," IBM J. R & D 18, March 1974.

13. F. Baskett, K.M. Chandy, F.G. Muntz, and F.G. Palacios, "Open, Closed and Mixed Networks of Queues with Different Classes of Customers," JACM 22, April 1975, pp. 248-260.

14. S. Karlin and H. M. Taylor, "A First Course in Stochastic Processes," 2nd edition, Academic Press. 1975.

15. M. A. Crane and D. L. Iglehart, "Simulating Stable Stochastic System, I: General Multiserver Queues," JACM 21, January 1974, pp. 103-113.

16. D. P. Gaver, "Analysis of Remote Terminal Backlogs under Heavy Demand Conditions," JACM 18, July 1971.

17. H. Kobayashi, Y. Onozato, and D. Huynch, "An Approximate Method for Design and Analysis of an ALOHA System," IEEE Trans. Commun., Com 25, January 1977, pp. 148-157.

18. J. F. C. Kingman, "Some Inequalities for the Queue in the GI/G/1," Biometrika 49, 1962.

19. J. Keilson, "The Role of Green's Functions in Congestion Theory," Proceedings of the Symposium on Congestion Theory, W. L. Smith and W. E. Wilkinson (ed) U. of North Carolina, 1965, pp. 43-71.

20. P. A. Lewis and G. S. Shedler, "Empirically Derived Micromodels for Sequences of Page Exceptions," IBM J. R & D 17, March 1973.

21. H. A. Anderson and R. Sargent, "The Statistical Evaluation of the Performance of an Experimental APL/360 System," Statistical Computer Performance Evaluation, W. Freiberger (ed.) Academic Press, 1972, pp. 73-98.

22. G. S. Shedler, "A Cyclic Queue Model of a Paging Machine," IBM Research Report RC-2814, 1970.

23. D. Iglehart, "Queueing Theory," preprint, O.R. dep. Stanford U.

24. T. Price, "A Note on the Effect of the Central Processor Service Time Distributuion on Processor Utilization in Multiprogrammed Computer Systems," JACM 25, April 1976.

25. B. Halachmi and W. R. Franta, "A Closed, Cyclic, Two Stage Multiprogrammed System Model and Its Diffusion Approximation Solution," Proc. SIGMETRIC Symposium, Montreal, October 1974; ACM Performance Evaluation Review 3, 1974, pp. 54-64.

26. B. Halachmi and W. R. Franta, "A Diffusion Approximate Solution to the G/G/K Queueing System," TR-75-3, Department of Computer Science, the University of Kansas.

27. W. R. Franta, "The Mathematical Analysis of the Computer System Modeled as a Two Stage Cyclic Queue," Acta Informatica 6, 1976, pp. 187-209.

28. L. Ponzin, "CIGALE, the Packet-Switching Machine of the Cyclades Computer Network," Inf. Proc. 74, Stockholm, North-Holland, August 1974.