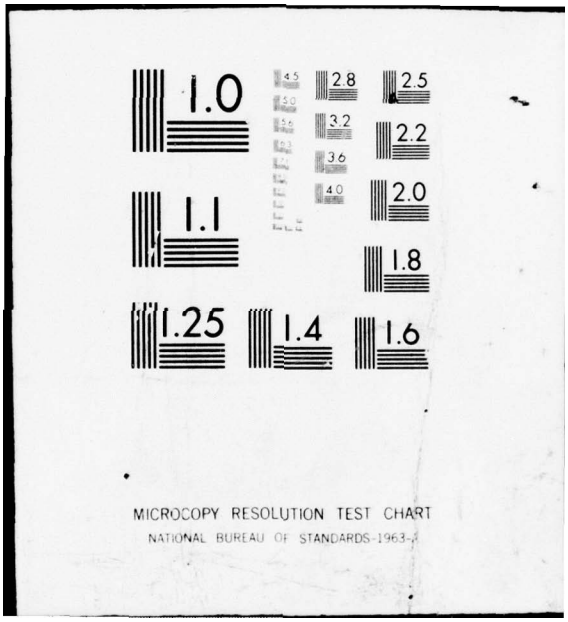
AD-A044 741 CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS F/G 9/2
MODELS FOR WORK BACKLOGS AT COMPUTERS THAT TIME-SHARE HETEROGEN--ETC(U)
JUL 77 J P LEHOCZKY, D P GAVER AFOSR-74-2642

UNCLASSIFIED TR-132 AFOSR-TR-77-1215 NL

| OF |
AD
A044 741

END
DATE
FILMED
10-77
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

AFOSR-TR- 77-1215

4

B.S.

DEPARTMENT

OF

STATISTICS

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is
approved for publication IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 77 - 1215	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): MODELS FOR WORK BACKLOGS AT COMPUTERS THAT TIME-SHARE HETEROGENEOUS USERS,	5. TYPE OF REPORT & PERIOD COVERED: Interim report
6. PERFORMING ORG. REPORT NUMBER: Tech Report No 132	
7. AUTHOR(s): John P. Lehoczky and Donald P. Gaver	8. CONTRACT OR GRANT NUMBER(s): AFOSR 74-2642
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Carnegie-Mellon University, Department of Statistics, Pittsburgh, PA 15213	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research/NM, Bolling AFB, Washington, DC 20332	12. REPORT DATE: July 1977
	13. NUMBER OF PAGES: 38
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): TR-132	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
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)
time-sharing computers, repairman models, multitype customers, queue discipline, diffusion approximation, state space reduction

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

A number of repairman-type models useful for describing and evaluating time-sharing computer systems with multitype users are presented and analyzed. Several approximation methods are introduced including one which allows for performance evaluation as a function of the queue discipline. Diffusion approximations are also considered. The accuracy of each of the approximation methods is assessed by numerical methods.

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

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

391 190

MODELS FOR WORK BACKLOGS AT COMPUTERS THAT TIME-SHARE HETEROGENEOUS USERS

by

John P. Lehoczky*
Carnegie-Mellon University

Donald P. Gaver**
U.S. Naval Postgraduate School

July, 1977

Technical Report No. 132

Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA. 15213

* J. P. Lehoczky acknowledges the support of the Air Force Office of Scientific Research at Carnegie-Mellon University, Grant No. AFOSR-74-2642b.

** D. P. Gaver acknowledges the research support of the Office of Naval Research at the Naval Postgraduate School.

MODELS FOR WORK BACKLOGS AT COMPUTERS THAT TIME-SHARE HETEROGENEOUS USERS

Donald P. Gaver*

John P. Lehoczky**

1. Introduction

The typical time-sharing computer system experiences demand for computing effort from users at a number of remote terminals. In an interactive manner a terminal user successively addresses the computer, waits until his request is processed, examines the result for a time (the "think time"), submits a subsequent request, and so the process repeats. Requests submitted join a queue at the computer, and the size of that queue, with the implied delays, depends upon the nature of the requests submitted as well as the number of terminals, the think time duration, and the speed and scheduling strategy of the computer.

If, as seems natural, the character of the requests varies both over time and from terminal to terminal, a Markovian model to predict total computer backlog will require an extremely large state space. In this paper we present and evaluate several numerical

*D. P. Gaver acknowledges the research support of the Office of Naval Research at the Naval Postgraduate School.

**J. P. Lehoczky acknowledges the support of the Air Force Office of Scientific Research at Carnegie-Mellon University, Grant No. AFOSR-74-2642b.

or algorithmic, and also approximate analytical, methods for modelling complex time-sharing systems. As will be seen, useful approximations may be made that render the problem computationally tractable by reducing the size of an otherwise intolerably large state space. It is of interest that an analytical approach via diffusion approximations, theoretically valid when the number of terminals is large, actually yields quite acceptable model agreement in many cases.

In the following sections we formulate the models to be described next. Model 1 provides a simple diffusion approximation for backlogs when jobs submitted from terminals are of a single type. For such a system transient behavior and dynamic performance of the backlog is available in simple form. This model is generalized in Model 2, wherein a mixture of two job types apply from all terminals, and are served in arrival order at the computer. Such a model requires specification of both the number of jobs enqueued and the identity of the job in service. Various approximations are employed to simplify this problem; in particular a time-weighted processor-sharing approximation proves to be effective; see Lehoczky and Gaver (1977). Analytical solutions follow when the latter approach is utilized in connection with a diffusion approximation. Finally, Model 3 segregates terminals into two types, each submitting jobs of a specific, but different, character. Again both numerical and analytical solutions are obtained, and their numerical agreement is shown to be satisfactory.

The models proposed should prove to be useful for planning and evaluating both new and existing time-sharing computer systems.

2. Model 1: Diffusion Approximations for One Job Type

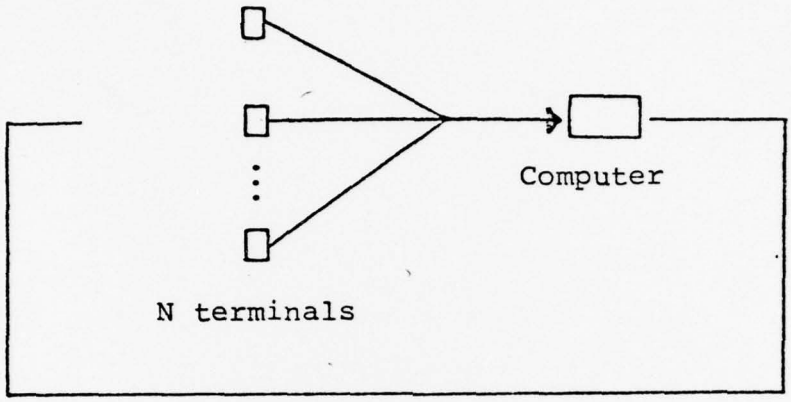


FIGURE 1

In this section we assume that each of N terminals is active and requests service according to an exponential distribution with parameter λ. Each job requires service from the computer and the service times also have an exponential distribution with parameter $N\mu$, independently of the arrivals. All jobs are identical, and the computer is assumed to operate according to a first-come, first-served (FCFS) queue discipline.

We define $Q(t)$ to be the number of jobs at the computer at time t. Clearly $\{Q(t), t \geq 0\}$ is a continuous-time Markov chain. One can explicitly solve the steady state birth-death equations (see Karlin (1967, p.208) for example). This model has been extensively studied as the "repairman" model; Feller (1957).

It is somewhat difficult to understand the dynamic or transient behavior of such a system. One could solve the Kolmogorov forward equations, perhaps in closed form by means of Laplace transforms, or perhaps numerically; however, an approach which gives a great deal of insight into the dynamics of the system is to introduce a diffusion approximation for $\{Q(t),\ t \geq 0\}$ We note that if the number of terminals N is large, changes in $Q(t)$ will occur frequently. Over any short period of time, the system size will change by the sum of many independent Bernoulli-like random variables. In the spirit of the central limit theorem, then, we model these changes by normal random variables. The resulting state space will be $\mathbb{R}^+ = [0,\infty)$ rather than $\{0, 1, \ldots, N\}$. A rigorous theory of this approximation has been supplied by Barbour (1974).

The process $\{Q(t),\ t \geq 0\}$ is thus approximated by a continuous-time Markov process satisfying the Ito type stochastic differential equation

$$dQ(t) = \lambda[N - Q(t)]dt - \mu N\, dt + \sqrt{\lambda(N - Q(t))}\, dW_1(t) - \sqrt{\mu N}\, dW_2(t) \qquad (2.1)$$

where $W_1(t)$ and $W_2(t)$ are independent standard Wiener processes. Actually, the parameters λ and μ may be time dependent. For what follows we assume λ and μ constant, and $\mu/\lambda < 1$. The argument for the model (2.1) is simply as

follows. The term $\lambda[N - Q(t)]$ represents the expected rate of arrivals when the number of users at the terminal is $Q(t)$, and μN represents the corresponding expected rate of departures; together they are called the drift. The term $\sqrt{\lambda[N-Q(t)]}\, dW_1(t)$ represents the random flucutation of the arrival component to Q's increase, and $\sqrt{\mu N}\, dW_2(t)$ represents the corresponding independent fluctuation in departures. Our approximation regards both of these fluctuations as independently Gaussian.

Next we define a new process $\{X_N(t),\ t \geq 0\}$:

$$X_N(t) = (Q(t) - Nq(t))/\sqrt{N} \quad \text{or} \quad Q(t) = Nq(t) + \sqrt{N}\, X_N(t)\ . \tag{2.2}$$

Using Ito's Lemma (Arnold, p. 90), one may derive the stochastic differential equation governing $\{X_N(t),\ t \geq 0\}$. We find

$$dX_N(t) = \sqrt{N}\left(\frac{dq(t)}{dt} - \lambda(1-q(t))\right)dt + \mu dt - \lambda X_N(t)dt$$

$$+ \sqrt{\lambda(1-q(t))}\, dW_1(t) + \sqrt{\mu}\, dW_2(t) \tag{2.3}$$

Now let $N \to +\infty$. For $\{X_N(t),\ t \geq 0\}$ to converge weakly to a finite limit $\{X(t),\ t \geq 0\}$, the coefficient of the $\sqrt{N}\, dt$ term must be identically zero for all t. This compels us to state that

$$q'(t) = \lambda(1-q(t)) - \mu \ , \qquad q(0) = \frac{Q(0)}{N} \ . \tag{2.4}$$

Equation (2.4) is easily solved to give

$$q(t) = (q(0) - (1-\mu/\lambda))e^{-\lambda t} + (1-\mu/\lambda), \qquad \mu < \lambda \tag{2.5}$$

When $\mu \geq \lambda$ this particular approximation has little value, for it predicts an eventual queue of zero length; this occurs because we have not accounted for the presence of a boundary at $Q = 0$, encountered with noticeable frequency only when the system is very lightly loaded. Note that in such a situation the appropriate diffusion approximation is that for a single-server queue with Poisson arrivals; see Gaver (1968). Here we model systems that are rather heavily loaded, and ignore the boundary.

It $q(t)$ is chosen as in (2.5), then the results of Kurtz (1971) and Barbour (1974) insure that $\{X_N(t), t \geq 0\}$ will converge weakly to $\{X(t), t \geq 0\}$, a diffusion process governed by the stochastic differential equation

$$dX(t) = -\lambda X(t)\ dt + \sqrt{\lambda(1-q(t)) + \mu}\ dW(t) \ , \qquad X(0) = 0. \tag{2.6}$$

Note that we have consolidated the two random fluctuation terms into one.

Equation (2.6) characterizes a non-stationary Ornstein-Uhlenbeck process. Many results are known for these processes (see Arnold, Chapter 8). The diffusion approximation of

$\{Q(t), t \geq 0\}$ amounts to approximating $Q(t)$ by $Nq(t) + \sqrt{N}\, X(t)$ rather than $Nq(t) + \sqrt{N}\, X_N(t)$. The quantity $Nq(t)$ is referred to as the deterministic approximation, while $\sqrt{N}\, X(t)$ is a stochastic noise process superimposed on the deterministic term.

One can use (2.5) and (2.6) to study the dynamic or transient behavior of the system. If we let $t \to \infty$, an account of the steady-state behavior is obtained. In all cases $Q(t)$ will be approximately normally distributed. The parameters of this distribution are easily characterized. If $v(t) = \mathrm{Var}(X(t))$, then (Arnold, Chapter 8)

$$E(X(t)) = 0$$

and (2.7)

$$v'(t) = -2\lambda(t) + \lambda(1-q(t)) + \mu, \qquad v(0) = 0$$

Equation (2.7) can be easily solved using (2.5) to give

$$v(t) = \frac{\mu}{\lambda}\,(1 - e^{-\lambda t}) + (1 - q(0))(e^{-\lambda t} - e^{-2\lambda t}). \tag{2.8}$$

The marginal distribution of $Q(t)$ is approximately normal with mean $Nq(t)$ and variance $Nv(t)$. The transient behavior can be determined using the above analysis. If at some time t_0, $Q(t_0) = Nq_0$, then $Q(t)$ will be approximately normally distributed with mean and variance given by (2.5) and (2.8) with t replaced by $t-t_0$ and $q(0)$ replaced by q_0.

As $t \to \infty$, steady state prevails. The distribution of $Q(t)$ converges weakly to a normal distribution with mean $N(1 - \mu/\lambda)$ and variance $N\mu/\lambda$.

Equivalent diffusion approximation ideas, but given in a much different form, were originally presented by Iglehart (1965).

3. Model 2: Two Types of Customers

Model I is clearly an oversimplified representation of a time-sharing computer system. The computer, considered as a server, actually performs complex functions, and appears to be a network of queues and servers. We will, throughout this paper, continue to model the computer as a single server with a first-come-first-served queue discipline. However, we will generalize the exponential service distribution to a wider class, say the class of phase distributions. A second problem arises with the assumption of exponential user "think times," which again may appear unrealistic. Finally, it is unreasonable to assume that all of the terminal users request the same type of service from the computer. In fact, some users may be performing simple editing tasks, others may be making heavy use of the computer for scientific computations, while others may be using it in a mixture of these ways. This type of situation with heterogeneous job types is often modelled by assuming that the effective service time is a fixed mixture of individual service distributions. Unfortunately, in this context such an assumption seems unsatisfactory, since the appropriate mixture must ultimately depend on the types of jobs in service and must therefore change dynamically.

In modelling computer systems with a mixture of job types, a processor sharing queue discipline is typically assumed. Processor sharing is the limiting version of a "round robin" queue discipline as the quantum of service given

to each job approaches zero. Processor sharing provides great mathematical convenience in that resulting queueing networks will, under weak assumptions, exhibit a product-form steady state distribution (see Baskett et al. (1975)). Unfortunately, processor sharing may not faithfully model the behavior of the computer. The computer scheduling policy is closer to a FCFS queueing system than to a processor sharing one, assuming that actual quanta are relatively large, and most jobs finish service before the end of a quantum. The models that we consider better portray a batch system with terminal users than a true time-sharing system. The central-server type of multiprogramming model will be considered in another paper.

In this section, we assume there are two types of jobs, each with an exponential service time, with parameters μ_1 and μ_2. In Model 2 we assume that each time a job is submitted from a terminal, there is a probability p of its being type 1, and 1-p of its being Type 2. We assume jobs arrive from any terminal independently according to an exponential (λ) distribution and that the computer handles jobs in a FCFS fashion.

This system can be simply modelled as a Markov chain. Ordinarily with two types of customers in a single server queue, one must keep track of the order of the jobs in the queue. In this case, however, customers need not have their

identities established until they enter service. Consequently, if $Q(t)$ represents the number of jobs in service at time t and $I(t) = i$ if a type i job is in service, then $\{(N(t), I(t)), t \geq 0\}$ is a continuous time Markov chain with transitions given by

Transition		Rate
$(M,1) \longrightarrow (M+1,1)$	$M \geq 0$	$\lambda(N-M)\ dt + o(dt)$
$(M-1,1)$	$M \geq 1$	$\mu_1 p\ dt + o(dt)$
$(M-1,2)$	$M \geq 1$	$\mu_1(1-p)\ dt + o(dt)$
$(M,2) \longrightarrow (M+1,2)$	$M \geq 0$	$\lambda(N-M)\ dt + o(dt)$
$(M-1,1)$	$M \geq 1$	$\mu_2 p\ dt + o(dt)$
$(M-1,2)$	$M \geq 1$	$\mu_2(1-p)\ dt + o(dt)$
$(0,1) \longrightarrow (1,1)$		$\lambda N p\ dt + o(dt)$
$(0,2) \longrightarrow (1,2)$		$\lambda N(1-p)\ dt + o(dt)$

(3.1)

The states $(0,1)$ and $(0,2)$ can be aggregated into a single state representing system emptiness.

The steady state equations can be solved, although not in a convenient closed form. A Gauss-Seidel iteration procedure will provide an efficient numerical approach even for very large N in view of the sparseness of the transition matrix.

We wish to study three approximation methods for this simple model. Based on their performance we may be able to use them in more complicated models. The first method is a straightforward generalization of Model 1 and is based heavily on the fact that any terminal can submit either type of job; an assumption requiring jobs to be typed only upon reaching service. To approximate the behavior of this system in the multitype setting, we can replace the two service rates μ_1 and μ_2 by a weighted rate $\mu^* = (p/\mu_1 + (1-p)/\mu_2)^{-1}$. This approximation is intuitive in that it gives the correct mean service rate and is easy to handle. One can apply the analysis given in Section 2 by merely replacing μ by μ^*. Numerical examples of the accuracy of this method will be given later. In general, it provides a high level of accuracy when compared with an exact solution using (3.1), but the reader should note that it will not be easily applied to situations in which the number of terminals of each type is fixed and each terminal submits jobs of a particular type only.

The second approximation is based upon a processor sharing queue discipline concept modified to represent FCFS system behavior. If the queue were to contain $N_i(t)$ jobs of type i at time i, i = 1, 2, then the traditional processor sharing model allocates $N_i(t)/(N_i(t) + N_2(t))$ time units to jobs of type i, i = 1, 2 during each $(t, t + dt)$ time interval. Each job waiting for service is given equal weight; consequently, short jobs are given preferential treatment,

for they are not blocked behind long jobs. For this queue discipline $\{(N_1(t), N_2(t)), t \geq 0\}$ is a continuous time Markov chain and, using the results of Baskett et al. (1975), the stationary distribution will be of a product form. A simple modification will allow us to model the behavior of a FCFS server. We weight each job by the average amount of time it requires for service. Thus if $\Delta_i(t) = N_i(t + dt) - N_i(t)$, $i = 1, 2$, then we assume

$$P(\Delta_1(t)=-1, \Delta_2(t)=0 \mid \underset{\sim}{N}(t)) = \frac{\mu_1 N_1(t)/\mu_1}{N_1(t)/\mu_1 + N_2(t)/\mu_2} dt + o(dt)$$

$$P(\Delta_1(t)=0, \Delta_2(t)=-1 \mid \underset{\sim}{N}(t)) = \frac{\mu_2 N_2(t)/\mu_2}{N_1(t)/\mu_1 + N_2(t)/\mu_2} dt + o(dt) \qquad (3.2)$$

The motivation for using these transition rates is as follows. Over a short period of time, say $(t, t + dt)$, at most one specific job has positive probability of being completed. This is in contrast to (3.2) in which either type has positive probability of completion. Nevertheless, if we focus on a longer period of time, say the amount of time needed to service all jobs currently in the queue, then on the average $N_i(t)/\mu_i$ time units will be spent servicing type i jobs, $i = 1,2$. Thus the server will spend approximately a fraction $(N_i(t)/\mu_i)/(N_1(t)/\mu_1 + N_2(t)/\mu_2)$ of the total time servicing type i customers. Approximating (3.1) by (3.2) amounts to applying this fraction, which changes dynamically,

to every interval of length dt rather than just to the longer interval in which all customers are served. This approximation is suggested only for steady state calculations, since it does not model the local behavior of a FCFS server faithfully.

The transitions (3.2) belong to a general parametric class of transitions of the form

$$P(\Delta_1(t)=-1,\ \Delta_2(t)=0 \mid \underset{\sim}{N}(t)) = \frac{\mu_1 N_1(t)}{N_1(t) + cN_2(t)}\, dt + o(dt)$$

$$P(\Delta_1(t)=0,\ \Delta_2(t)=-1 \mid \underset{\sim}{N}(t)) = \frac{\mu_2 cN_2(t)}{N_1(t) + cN_2(t)}\, dt + o(dt) \qquad (3.3)$$

The parameter c governs the relative weight given to jobs of either type. If $c = 1$, then each is given equal weight, which corresponds to a processor sharing. If $c = \mu_1/\mu_2$, then the queue discipline becomes FCFS. As $c \to +\infty$, type 2 jobs receive priority, while if $c \to 0$, type 1 jobs receive priority. Equations (3.3) provide a new approach to assessing system performance as a function of the scheduling or queueing disciplines.

Equations (3.3) coupled with the usual equations governing arrivals to the queue in (3.1) provide the transition rates for $\{(N_1(t), N_2(t)),\ t \geq 0\}$. These equations can be solved numerically using Gauss-Seidel iteration. The accuracy of this method and selected numerical results will be given later.

The third modelling approach utilizes a diffusion approximation based on (3.3). This method, while perhaps less accurate than the first two has the virtue of providing simple closed-form expressions for various system parameters such as the expected queue lengths. It is difficult to make such assessments using numerical methods.

We follow the method outlined in Section 2 by writing stochastic differential equations to approximate $N_i(t)$, $i = 1,2$. Hence

$$\begin{aligned} dN_1(t) &= \lambda p(N-N_1(t)-N_2(t))dt - \mu_1 N \frac{N_1(t)}{N_1(t)+cN_2(t)}\, dt \\ &\quad + \sqrt{\lambda p(N-N_1(t)-N_2(t))}\, dW_1(t) \\ &\quad - \sqrt{\mu_1 N \frac{N_1(t)}{N_1(t)+cN_2(t)}}\, dW_2(t) \end{aligned} \tag{3.4}$$

$$\begin{aligned} dN_2(t) &= \lambda(1-p)(N-N_1(t)-N_2(t))dt - \mu_2 N \frac{N_1(t)}{N_1(t)+cN_2(t)}\, dt \\ &\quad + \sqrt{\lambda(1-p)(N-N_1(t)-N_2(t))}\, dW_3(t) \\ &\quad - \sqrt{\mu_2 N \frac{cN_1(t)}{N_1(t)+cN_2(t)}}\, dW_4(t) \end{aligned}$$

We next define $X_{iN}(t) = (N_i(t)-Nq_i(t))/\sqrt{N}$, $i = 1,2$ so that $(N_1(t),N_2(t)) = N(q_1(t),q_2(t)) + \sqrt{N}\,(X_{1N}(t),X_{2N}(t))$.

The first term provides the deterministic approximation, while the second is the stochastic noise term. To calculate representations for $q_i(t)$ and $X_{iN}(t)$, $i = 1,2$, we need an asymptotic expansion for $N_i(t)/(N_1(t)+N_2(t))$,

$$\frac{N_1(t)}{N_1(t)+cN_2(t)} = \frac{q_1(t)}{q_1(t)+cq_2(t)} + \frac{c}{\sqrt{N}}\left(\frac{q_2(t)X_{1N}(t)-q_1(t)X_{2N}(t)}{(q_1(t)+cq_2(t))^2}\right)+o(\frac{1}{\sqrt{N}})$$

$$\frac{cN_2(t)}{N_1(t)+cN_2(t)} = \frac{cq_2(t)}{q_1(t)+cq_2(t)} + \frac{c}{\sqrt{N}}\left(\frac{q_1(t)X_{2N}(t)-q_2(t)X_{1N}(t)}{(q_1(t)+cq_2(t))^2}\right)+o(\frac{1}{\sqrt{N}}) \qquad (3.5)$$

We now let $N \to +\infty$. Using the methods outlined in Section 2, $\{(X_{1N}(t),X_{2N}(t)),\ t \geq 0\}$ converges weakly to $\{(X_1(t),X_2(t),\ t \geq 0\}$ provided $(q_1(t),\ q_2(t))$ satisfies a system of ordinary differential equations. We find using (3.4) and (3.5) that

$$q_1'(t) = \lambda p(1-q_1(t)-q_2(t)) - \frac{\mu_1 q_1(t)}{q_1(t)+cq_2(t)}$$

$$q_2'(t) = \lambda(1-p)(1-q_1(t)-q_2(t)) - \frac{\mu_2 cq_2(t)}{q_1(t)+cq_2(t)} \qquad (3.6)$$

with $(q_1(0),q_2(0)) = (N_1(0),N_2(0))/N$.

The noise process $\{(X_1(t),X_2(t)),\ t \geq 0\}$ will satisfy

$$d\underset{\sim}{X}(t) = \underset{\sim}{A}_t \underset{\sim}{X}_t dt + \underset{\sim}{B}_t d\underset{\sim}{W}_t , \qquad \underset{\sim}{X}(0) = \underset{\sim}{0} \quad \text{a.s.} \tag{3.7}$$

where $\underset{\sim}{X}(t) = (X_1(t), X_2(t))^T$

$$\underset{\sim}{A}_t = \begin{pmatrix} -\lambda p - \dfrac{\mu_1 c q_2(t)}{(q_1(t)+cq_2(t))^2} & -\lambda p + \dfrac{\mu_1 c q_1(t)}{(q_1(t)+cq_2(t))^2} \\ -\lambda(1-p) + \dfrac{\mu_2 c q_2(t)}{(q_1(t)+cq_2(t))^2} & -\lambda(1-p) - \dfrac{\mu_2 c q_1(t)}{(q_1(t)+cq_2(t))^2} \end{pmatrix}$$

and

$$\underset{\sim}{B}_t = \begin{pmatrix} \sqrt{\lambda p(1-q_1(t)-q_2(t))} & \sqrt{\dfrac{\mu_1 q_1(t)}{q_1(t)+cq_2(t)}} & 0 & 0 \\ 0 & 0 & \sqrt{\lambda(1-p)(1-q_1(t)-q_2(t))} & \sqrt{\dfrac{\mu_2 c q_2(t)}{q_1(t)+cq_2(t)}} \end{pmatrix}$$

Equation (3.7) defines a bivariate, nonstationary Ornstein-Uhlenbeck process. From Arnold (Chapter 8), we know $E(\underset{\sim}{X}(t)) = \underset{\sim}{0}$ for all t, and $\underset{\sim}{X}(t)$ will have a bivariate normal distribution with covariance matrix $\underset{\sim}{\Sigma}_t = E(\underset{\sim}{X}(t)\ \underset{\sim}{X}^T(t))$ given by the unique non-negative-definite solution of

$$\dot{\underset{\sim}{\Sigma}}_t = \underset{\sim}{A}_t\underset{\sim}{\Sigma}_t + \underset{\sim}{\Sigma}_t\underset{\sim}{A}_t^T + \underset{\sim}{B}_t\underset{\sim}{B}_t^T , \qquad \underset{\sim}{\Sigma}_0 = 0 \tag{3.8}$$

This matrix Riccati equation can be solved numerically (see Arnold, Chapter 8 for a general expression).

The mean of $\underset{\sim}{N}(t)$ is $N(q_1(t),q_2(t))$ and may be found by solving (3.6). Because (3.6) is non-linear, it will have to be solved numerically. In fact, we will not now be concerned with the dynamic behavior predicted by the models (3.6) and (3.7). Instead, we choose to examine the predicted steady-state behavior that is obtained by letting $t \to +\infty$. As $t \to \infty$, $q_i(t) \to q_i$ and $A_t \to A$, $B_t \to B$. We find

$$\begin{aligned} q_1 &= (1 - 1/(\rho_1 + \rho_2))/(1 + \rho_2/c\rho_1) \\ q_2 &= (1 - 1/(\rho_1 + \rho_2))\rho_2/c\rho_1(1 + \rho_2/c\rho_1) \end{aligned} \tag{3.9}$$

where $\rho_i = \lambda p/\mu_i$, $i = 1,2$.

$\underset{\sim}{N}(t)$ will be approximately bivariate normally distributed with mean $N(q_1,q_2)$ and covariance matrix $N\underset{\sim}{\Sigma}$ where $\underset{\sim}{\Sigma}$ is the unique non-negative definite solution of

$$\underset{\sim}{A}\underset{\sim}{\Sigma} + \underset{\sim}{\Sigma}\underset{\sim}{A}^T = -\underset{\sim}{B}\underset{\sim}{B}^T \tag{3.10}$$

with

$$\underset{\sim}{A} = \begin{pmatrix} -\lambda p - \dfrac{\mu_1 c q_2}{(q_1+cq_2)^2} & -\lambda p + \dfrac{\mu_1 c q_2}{(q_1+cq_2)^2} \\ -\lambda(1-p) + \dfrac{\mu_2 c q_2}{(q_1+cq_2)^2} & -\lambda(1-p) - \dfrac{\mu_2 c q_1}{(q_1+cq_2)^2} \end{pmatrix}$$

and

$$\underset{\sim}{B}\underset{\sim}{B}^T = \begin{pmatrix} \lambda p(1-q_1-q_2) + \dfrac{\mu_1 q_1}{q_1+cq_2} & 0 \\ 0 & \lambda(1-p)(1-q_1-q_2) + \dfrac{\mu_2 c q_2}{q_1+cq_2} \end{pmatrix}$$

One use of (3.9) and (3.10) is to study the influence of the parameter c on system performance. If $c = \rho_2/\rho_1$, then the values of q_1 and q_2 are equal. If $c < \rho_2/\rho_1$, then $q_2 > q_1$, whereas $q_2 < q_1$ if $c > \rho_2/\rho_1$. It should be noted that no matter what c is chosen, $q_1 + q_2$ will equal $1 - 1/(\rho_1 + \rho_2)$. This is intuitive, since $q_1 + q_2$ represents the number of jobs in service. No matter which value of c is chosen, the same backlog of jobs will be faced by the server. The parameter c controls the relative fraction of jobs in the queue of each type and thus the response time for each job type. Here $q_1/q_2 = c\rho_1/\rho_2$. If costs are assigned to waiting or turnaround time, then an appropriate value of c

could be chosen to minimize expected cost. This value of c can then be implemented by using the corresponding scheduling algorithms.

4. Numerical Results for Model 2

In this section we present numerical results to assess the accuracy of the three approximating methods given in Section 3. The methods referred to in the following tables are:

Method 1: Exact solution of (3.1) using Gauss-Seidel iteration.

Method 2: Calculations made using $\mu^* = 1/(p/\mu_1 + (1-p)/\mu_2)$ and assuming the number of Type 1 jobs in line has a binomial distribution.

Method 3: Weighted processor-sharing system (3.2) using Gauss-Seidel iteration.

Method 4: Diffusion approximation using (3.2), (3.9), and (3.10).

Here are numerical results for particular cases to which these methods have been applied.

Case 1. $N = 10$, $p = 1/2$, $\lambda = 10$, $N\mu_1 = 10$, $N\mu_2 = 20$ ($\therefore$ $\rho_1 = 1/2$, $\rho_2 = 1/4$).

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$*
1	4.50	4.17	2.58	2.41	-.72
2	4.50	4.17	2.47	2.47	-.73
3	4.53	4.14	2.66	2.29	-.73
4	4.33	4.33	2.70	2.31	-.73

*Here $\rho(N_1,N_2)$ refers to the ordinary product-moment correlation between N_1 and N_2.

Case 2. $N = 10$, $p = 1/4$, $\lambda = 10$, $N\mu_1 = 10$, $N\mu_2 = 20$
($\therefore$ $\rho_1 = 1/4$, $\rho_2 = 3/8$).

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	2.25	6.15	1.79	2.43	-.62
2	2.25	6.15	1.73	2.53	-.63
3	2.28	6.12	1.84	2.26	-.61
4	2.10	6.30	2.07	2.82	-.69

Case 3. $N = 10$, $p = 1/2$, $\lambda = 10$, $N\mu_1 = 10$, $N\mu_2 = 50$
($\therefore$ $\rho_1 = 1/2$, $\rho_2 = 1/10$).

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	4.50	3.83	2.67	2.39	-.56
2	4.50	3.83	2.39	2.39	-.64
3	4.58	3.76	2.83	1.98	-.63
4	4.17	4.17	2.99	2.09	-.66

Case 4. $N = 10$, $p = 3/4$, $\lambda = 10$, $N\mu_1 = 10$, $N\mu_2 = 20$
($\therefore$ $\rho_1 = 3/4$, $\rho_2 = 1/8$).

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	6.75	2.11	2.33	1.65	-.71
2	6.75	2.11	2.24	1.67	-.73
3	6.76	2.10	2.40	1.61	-.73
4	6.64	2.21	1.98	1.16	-.63

These four cases were chosen to be quite extreme in order to provide a reasonable test of the various methods. We have also provided the variances of N_1 and N_2 rather than the standard deviations; the latter would exhibit small percentage differences than do the variances.

For $N = 10$, there would seem to be little hope that Method 4, the diffusion approximation, will be very accurate. This method requires $N \to +\infty$, yet even for $N = 10$ the accuracy is surprising especially for the means. In all cases $E(N_1) + E(N_2)$ is very nearly exact, and the individual components are off by at most 1/2. This error arises from the fact that the different types have different amounts of service time and hence different server occupancies, but this discrepancy decreases as N increases. The variances are less satisfactory in terms of percentage error, but the absolute errors are probably small enough to be practically negligible. Accuracy increases with N, and the method is appealing because of the closed form expressions produced.

Method 2 offers a great deal of accuracy in most cases. Unfortunately, it is very special and is not easily extended to more complex models, in particular to those with a fixed number of terminals of each type. In such a case one must keep track of the order of jobs in the queue. Even changing from an exponential to a general phase type distribution will reduce the accuracy of Method 2; the accuracy of variance estimates will suffer more than that of mean estimates.

Method 3 gives acceptable results, better than Method 4 but not as good as Method 2. Nevertheless, this method offers great promise in that it can be generalized to handle more complicated systems and phase-type service distributions. The approximation given in (3.3) for two types of service can be extended to many types and phase type service. Furthermore, it offers the way to studying the performance of the system as a function of the queue discipline. If N is large and the service rates are not widely different, the accuracy of Method 2 should be accurate enough for most purposes.

5. Model 3: Two Types of Terminals

In this section, we remove the assumption that each terminal can submit either type of job, a fraction p of which are of Type 1. Instead, we assume that each terminal submits only jobs of a certain fixed type. Specifically, let M_i terminals submit only jobs of type i, i = 1,2, and let $N = M_1 + M_2$ be the total number of terminals. Jobs of type i require an exponential (μ_i) processing time. Later we will generalize to allow phase type service distributions.

The "fixed terminal" assumption is not ideal but it probably represents the real situation more faithfully than does the assumption that terminals can submit any job type at any time. It seems plausible that terminal users typically interact with a single problem type, hence job type, for a significant period of time. As a result, the fixed terminal assumption is likely to be reasonable, if not ideal, over a moderate period of time. Modifications which allow for the sign-on and sign-off of users, changes in job types, and the influence of queue discipline on such a system will be considered in a subsequent paper.

We let $N_i(t)$ denote the number of type i jobs in service at time t. The arrival rate of type i jobs to the computer is $\lambda_i(M_i - N_i(t))dt + o(dt)$, i = 1.2.

An exact analysis of this stochastic model is very difficult because the order of customers in a FCFS queue must be kept as a part of the state description. The pair $(N_1(t),N_2(t))$ alone is not sufficient to guarantee the Markov property. A Markovian

state description will require a very large state space, and any solution must apparently be computed numerically. For phase-type service distributions and M_1 and M_2 around 25, it seems that one must resort to simulation. As a result, it is difficult to assess the dependence of system performance variables such as queue lengths and utilization on service rates, input rates, number of terminals, and queue discipline. Consequently, we seek approximate solutions which lead to tractable results. We present three such methods. First, the weighted processor sharing FCFS approach introduced in Section 3 can be adopted. This will not lead to closed form expressions but will considerably reduce the state space. The state space will be of the order of $M_1 \cdot M_2$ for exponential service times, but of course much larger for phase-type service distributions. Second, we give a diffusion approximation for the system based on the weighted processor sharing method. Third, a new method which keeps exact track of the job in service will be introduced. The third method has the advantage of a small state space even for general phase distributions and is more accurate than the other two. Even so, it is designed to handle only FCFS queues and does not seem to lead to a simple diffusion approximation with associated closed-form expressions.

The weighted processor sharing FCFS approximation is defined by the following transitions and transition rates.

Transitions	Rate
$(N_1\ N_2) \longrightarrow (N_1-1, N_2)$	$\frac{\mu_1 N_1}{N_1 + cN_2}\, dt + o(dt)$
(N_1, N_2-1)	$\frac{\mu_2 cN_2}{N_1 + cN_2}\, dt + o(dt)$
(N_1+1, N_2)	$\lambda_1(1-N_1)dt + o(dt)$
(N_1, N_2+1)	$\lambda_2(M_2-N_2)dt + o(dt)$.

(5.1)

One can derive balance equations and solve them numerically for given values of the parameters λ_j and μ_j.

Phase-type distributions can be easily incorporated by defining the number of type i jobs currently in phase j of their service. The rate at which such jobs complete service is taken to be $\mu_{ij}c_{ij}N_{ij}(t)/(\sum_k\sum_\ell c_{k\ell}N_{k\ell}(t))$. The constants $\{c_{k\ell}\}$ can be chosen to model various types of queue disciplines. Generally, letting $c_{ij} = 1/\mu_{ij}$ will give FCFS, $c_{ij} = 1$ will give processor sharing, and c_{ij} close to zero or infinity represents various priority systems.

To obtain more insight into (5.1), we next develop a diffusion approximation following the methods outlined in Sections 2 and 3. Specifically, we treat $N_1(t)$ and $N_2(t)$ as continuous variables

whose probabilistic structure is given by the stochastic differential equations

$$dN_1(t) = \lambda_1(M_1-N_1(t))dt - \frac{\mu_1 N\, N_1(t)}{N_1(t) + cN_2(t)}\, dt$$

$$+ \sqrt{\lambda_1(M_1-N_1(t))}\, dW_1(t) - \sqrt{\frac{\mu_1 N\, N_1(t)}{N_1(t) + cN_2(t)}}\, dW_2(t) \tag{5.2}$$

$$dN_2(t) = \lambda_2(M_2-N_2(t))dt - \frac{\mu_2 N\, cN_2(t)}{N_1(t) + cN_2(t)}\, dt$$

$$+ \sqrt{\lambda_2(M_2-N_2(t))}\, dW_3(t) - \sqrt{\frac{\mu_2 N\, cN_2(t)}{N_1(t) + cN_2(t)}}\, dW_4(t).$$

We next define $X_{iN}(t) = (N_i(t) - Nx_i(t))/\sqrt{N}$, $i = 1,2$ and let $\alpha_i = N_i(t)/N$, $i = 1,2$. Again, if $(x_1(t), x_2(t))$ satisfies a certain set of ordinary differential equations, then as $N \to \infty$ $\{(X_{1N}(t), X_{2N}(t), t \geq 0\}$ converges weakly to a limiting stochastic process $\{(X_1(t), X_2(t)), t \geq 0\}$. The deterministic approximation is given by $N(x_1(t), x_2(t))$ where

$$\begin{aligned} x_1'(t) &= \lambda_1(\alpha_1-x_1(t)) - \mu_1 x_1(t)/(x_1(t) + cx_2(t)) \\ x_2'(t) &= \lambda_2(\alpha_2-x_2(t)) - \mu_2 cx_2(t)/(x_1(t) + cx_2(t)) \end{aligned} \tag{5.3}$$

with $x_i(0) = N_i(0)/N$. The noise process $\{(X_1(t), X_2(t)), t \geq 0\}$ will, according to Ito's Lemma, satisfy

$$d\underset{\sim}{X}(t) = \underset{\sim t}{A}\underset{\sim}{X}(t)dt + \underset{\sim t}{B}d\underset{\sim}{W}(t) , \qquad \underset{\sim}{X}(0) = \underset{\sim}{0} \tag{5.4}$$

where $\underset{\sim}{X}(t) = (X_1(t), X_2(t))^T$,

$$\underset{\sim t}{A} = \begin{pmatrix} -(\lambda_1 + c\mu_1 x_2/(x_1+cx_2)^2) & + cx_1\mu_1/(x_1+cx_2)^2 \\ + c\mu_2 x_2/(x_1+cx_2)^2 & -(\lambda_2 + c\mu_2 x_1/(x_1+cx_2)^2) \end{pmatrix}$$

and

$$\underset{\sim t}{B} = \begin{pmatrix} \sqrt{\lambda_1(\alpha_1-x_1)} & -\sqrt{\dfrac{\mu_1 x_1}{x_1+cx_2}} & 0 & 0 \\ 0 & 0 & \sqrt{\lambda_2(\alpha_2-x_2)} & -\sqrt{\dfrac{\mu_2 cx_2}{x_1+cx_2}} \end{pmatrix}$$

The t-dependence of x_1 and x_2 has been suppressed in $\underset{\sim t}{A}$ and $\underset{\sim t}{B}$.

Equation (5.5) characterizes a nonstationary bivariate Ornstein-Uhlenbeck process with mean $\underset{\sim}{0}$. The covariance matrix of $(N_1(t),N_2(t))$ is thus approximately given by $N\underset{\sim t}{\Sigma}$ where $\underset{\sim t}{\Sigma}$ is the unique non-negative definite solution of (3.8). The mean of $(N_1(t),N_2(t))$ is approximately $N(x_1(t),x_2(t))$.

The steady-state behavior of the system is deduced by letting $t \to \infty$. The steady state means, (x_1, x_2), will satisfy

$$0 = \lambda_1(\alpha_1 - x_1) - \mu_1 x_1/(x_1 + cx_2)$$

$$0 = \lambda_2(\alpha_2 - x_2) - \mu_2 cx_2/(x_1 + cx_2) \quad , \tag{5.5}$$

while the covariance matrix, $N\underset{\sim}{\Sigma}_t$, can be derived from (3.10).

This analysis is presented only for exponential service, but can be easily extended to allow for general phase-type service distributions. Numerical examples will be presented in the next section.

The third approximation method offers excellent accuracy. In this method the state description keeps track of the number of jobs of each type awaiting service, as well as the phase of the current job receiving service. The exact order of jobs awaiting service is not recorded. Instead when a customer finishes service, the next customer is chosen at random from those waiting for service. For exponential service, the state of the system is given by $\{(N_1(t), N_2(t), I(t)), t \geq 0\}$ where $N_i(t)$ represents the number of type i jobs awaiting service and $I(t)$ gives the identity of the job in service. Then the transitions are given by

Transition		Rate
$(N_1,N_2,1)$	$\longrightarrow (N_1+1,N_2,1)$	$\lambda_1(M_1-N_1)dt + o(dt)$
	$(N_1,N_2+1,1)$	$\lambda_1(M_2-N_2)dt + o(dt)$
	$(N_1-1,N_2,1)$	$\mu_1 N(N_1/(N_1+N_2))dt + o(dt)$
	$(N_1,N_2-1,2)$	$\mu_1 N(N_2/(N_1+N_2))dt + o(dt)$
$(N_1,N_2,2)$	$\longrightarrow (N_1+1,N_2,2)$	$\lambda_1(M_1-N_1)dt + o(dt)$
	$(N_1,N_2+1,2)$	$\lambda_2(M_2-N_2)dt + o(dt)$
	$(N_1-1,N_2,1)$	$\mu_2 N(N_1/(N_1+N_2))dt + o(dt)$
	$(N_1,N_2-1,2)$	$\mu_2 N(N_2/(N_1+N_2))dt + o(dt)$.

(5.6)

This approximation can be easily extended to allow for phase-type service distributions, but apparently does not lend itself to closed-form approximations.

6. Numerical Results for Model 3.

In this section, we present numerical results for assessing the accuracy of the three approximation methods given in Section 5. The methods referred to in the following tables are:

Method 1: Solution of the exact system using simulation.

Method 2: Approximation based on (5.8) using simulation.

Method 3: Approximation based on (5,1)-weighted processor sharing with $c = \mu_1/\mu_2$ using simulation.

Method 4: Diffusion approximation using (5.5) and (3.10).

For the first three methods, a simulation consisting of 600,000 system transitions blocked into 20 groups of 30,000 was performed. The estimated standard deviation is given in parentheses. All cases consisted of Gamma $(2,N\mu_i)$ service times with $\mu_1 = 1$, $\mu_2 = 1/2$ and $M_1 = M_2 = 20$.

Case 1. $\lambda_1 = 2$, $\lambda_2 = 2$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	16.64 (.05)	16.67 (.05)	3.17 (.15)	2.46 (.07)	.02 (.03)
2	16.65 (.07)	16.65 (.04)	3.19 (.10)	2.47 (.06)	.02 (.02)
3	16.66 (.05)	16.67 (.09)	3.20 (.14)	3.20 (.14)	.12 (.03)
4	16.67	16.67	2.97	3.14	.09

Case 2. $\lambda_1 = 2, \lambda_2 = 1$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	16.23 (.06)	13.72 (.10)	3.53 (.13)	4.22 (.15)	.06 (.03)
2	16.25 (.08)	13.72 (.10)	3.54 (.18)	4.15 (.17)	.07 (.02)
3	16.26 (.06)	13.74 (.11)	3.46 (.11)	5.41 (.04)	.19 (.04)
4	16.28	13.72	3.36	5.45	.16

Case 3. $\lambda_1 = 1, \lambda_2 = 2$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	14.01 (.08)	16.49 (.05)	4.77 (.18)	2.57 (.08)	.06 (.02)
2	14.02 (.10)	16.48 (.04)	4.91 (.23)	2.55 (.09)	.05 (.02)
3	14.03 (.09)	16.49 (.06)	4.75 (.15)	3.37 (.12)	.14 (.03)
4	14.03	16.49	4.69	3.33	.11

Case 4. $\lambda_1 = 1, \lambda_2 = 1$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	13.31 (.08)	13.34 (.10)	5.35 (.18)	5.41 (.22)	.12 (.03)
2	13.28 (.12)	13.33 (.09)	5.53 (.27)	4.46 (.24)	.11 (.03)
3	13.31 (.10)	13.36 (.11)	5.38 (.19)	5.87 (.30)	.23 (.03)
4	13.33	13.33	5.28	5.88	.20

Case 5. $\lambda_1 = 1, \lambda_2 = 1/2$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	11.60 (.12)	8.33 (.12)	6.74 (.30)	6.13 (.35)	.26 (.03)
2	11.63 (.12)	8.35 (.14)	6.87 (.40)	6.05 (.35)	.27 (.03)
3	11.62 (.14)	8.41 (.19)	7.37 (.31)	7.64 (.55)	.40 (.04)
4	11.72	8.28	7.02	8.12	.37

Case 6. $\lambda_1 = 1/2, \lambda_2 = 1$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	9.27 (.10)	12.68 (.11)	6.27 (.22)	5.04 (.23)	.20 (.02)
2	9.28 (.13)	12.68 (.10)	6.37 (.31)	4.90 (.22)	.19 (.02)
3	9.26 (.14)	12.70 (.12)	6.46 (.31)	6.67 (.33)	.27 (.03)
4	9.28	12.68	6.36	6.64	.25

Case 7. $\lambda_1 = 1/2, \lambda_2 = 1/2$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	6.60 (.18)	6.74 (.14)	7.57 (.33)	7.11 (.31)	.42 (.03)
2	6.62 (.21)	6.75 (.20)	7.67 (.57)	7.11 (.50)	.43 (.04)
3	6.59 (.16)	6.82 (.17)	8.29 (.43)	9.00 (.54)	.52 (.03)
4	6.67	6.67	8.40	9.88	.52

Case 8. $\lambda_1 = 1/4$, $\lambda_2 = 1/2$

Method	$E(N_1)$	$E(N_2)$	$Var(N_1)$	$Var(N_2)$	$\rho(N_1,N_2)$
1	2.71 (.11)	4.86 (.17)	4.05 (.22)	7.31 (.37)	.49 (.02)
2	2.69 (.14)	4.81 (.19)	4.12 (.22)	7.31 (.32)	.50 (.02)
3	2.76 (.12)	5.09 (.15)	4.28 (.34)	9.12 (.34)	.54 (.02)
4	2.46	4.38	5.28	13.04	.66

The numerical examples, which cover a wide range of traffic intensities, illustrate the extraordinary agreement of the three methods in estimating the means. Even the diffusion approximation offers nearly exact answers. For the variances, Method 2, which keeps track of the customer type in service, appears to be nearly exact. The differences between 1 and 2 are dominated by the quoted error of simulation. On the other hand, Methods 3 and 4 do not provide good accuracy, especially for $Var(N_2)$ and $\rho(N_1,N_2)$. It is perhaps surprising that the diffusion approximation (Method 4) agrees so well with Method 3 in view of the small value of N ($N = 40$). It is clear that for phase-type distributions, the variances reported by the weighted processor sharing FCFS approximation will be inaccurate, providing only a rough estimate of the true value. On the other hand, the means are exceptionally good, and the diffusion approximation means provide trustworthy expressions for average system occupancy of each type.

We have not shown here the computer idleness probability. It behaves similarly:

Method 2 provides nearly an exact result while

Method 3 gives about 10 to 20 percent error.

The diffusion approximation applied here does not recognize idleness of the server.

REFERENCES

Arnold, L. (1974). Stochastic Differential Equations: Theory and Applications. Wiley-Interscience. John Wiley and Sons, New York, N.Y.

Barbour, A. (1974). "On a functional limit theorem for Markov population processes." Adv. in Appl. Prob., 6, pp. 21-29.

Baskett, F., Chandy, Muntz, R.R., and Palacios-Gomez, F. (1975). "Open, closed and mixed networks of queues with different classes of customers." JACM 22, pp. 248-260.

Feller, W. (1957). An Introduction to Probability Theory and Its Applications, I. John Wiley and Sons, New York.

Gaver, D. P. (1968). "Diffusion approximations and models for certain congestion problems." J. Appl. Prob. 5, pp. 607-623.

Gaver, D.P. and Lehoczky, J.P. (1977). "Approximate models for central server systems with two job types." Naval Postgraduate School Technical Report (in preparation).

Iglehart, D.L. (1965). "Limiting diffusion approximation for the many server queue and the repairman problem." J. Appl. Prob. 2, pp. 429-441.

Karlin, S. (1966). A First Course in Stochastic Processes, Academic Press.

Kurtz, T.G. (1971). "Limit theorems for sequence of jump Markov processes approximating ordinary differential equations." J. Appl. Prob. 8, pp. 344-356.