STAN-CS-73-360 SU-SEL-73-016

Open, Closed, and Mixed Networks of Queues With Different Classes of Customers

by

Richard R. Muntz and Forest Baskett

August 1972

Technical Report No. 33

This work was supported by Joint Services Electronics Programs: U.S. Army, U.S. Navy, U.S. Air Force under Contract N-00014-67-A-01 12-0044 and by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC-15-69-C-0258.

DIGITAL SYSTEMS LABORATORY

STANFORD ELECTRONICS LABORATORIES

STANFORD UNIVERSITY • STANFORD, CALIFORNIA

STAN-CS-73-360 SEL-73-016

OPEN, CLOSED, AND MIXED NETWORKS OF QUEUES WITH DIFFERENT CLASSES OF CUSTOMERS

by

Richard R. Muntz† and Forest Baskett

August 1972

Technical Report no. 33

DIGITAL SYSTEMS LABORATORY

Department of Electrical Engineering Department of Computer Science

Stanford University

Stanford, California

† Department of Computer Science, University of California at Los Angeles, Los Angeles, California.

This work was supported by Joint Services Electronics Programs: U.S. Army, U.S. Navy, U.S. Air Force under contract N-00014-67-A-0112-0044 and by the Advanced Research Projects Agency of the Department of Defense under contract DAHC-15-69-C-0258.

Open, Closed, and Mixed Networks of Queues with Different Classes of Customers

Richard R. Muntz
University of California at Los Angeles
Los Angeles, California

Forest Baskett
Stanford University
Stanford, California

ABSTRACT

We derive the joint equilibrium distribution of queue sizes in a network of queues containing N service centers and R classes of customers. The equilibrium state probabilities have the general form:

$$P(S) = C d(S) f_1(x_1) f_2(x_2) \ldots f_N(x_N)$$

where S is the state of the system, x_i is the configuration of customers at the ith service center, d(S) is a function of the state of the model, f_i is a function that depends on the type of the ith service center, and C is a normalizing constant. We consider four types of service centers to model central processors, data channels, terminals, and routing delays. The queueing disciplines associated with these service centers include first-come-first-served, processor sharing, no queueing, and last-come-first-served. Each customer belongs to a single class of customers while awaiting or receiving service at a service center but may change classes and service centers according to fixed probabilities at the completion of a service request. For open networks we consider state dependent arrival processes. Closed networks are those with no arrivals. A network may be closed with respect to some classes of customers and open with respect to other classes of customers. At three of

the four types of service centers, the service times of customers are governed by probability distributions having rational Laplace transforms, different classes of customers having different distributions. At first-come-first-served type service centers the service time distribution must be identical and exponential for all classes of customers. Many of the network results of Jackson on arrival and service rate dependencies, of Posner and Bernholtz on different classes of customers, and of Chandy on different types of service centers are combined and extended in this paper. The results become special cases of the model presented here. An example shows how different classes of customers can affect models of computer systems.

Finally, we show that an equivalent model encompassing all of the results involves only classes of customers with identical exponentially distributed service times. All of the other structure of the first model can be absorbed into the fixed probabilities governing the change of class and change of service center of each class of customers.

TABLE OF CONTENTS

Introduction

Networks of queues are important models of multiprogrammed and time-shared computer systems. Work on this application in the last several years has produced a variety of models meant to capture important aspects of computer systems. The results of this paper unify and extend a number of those separate results in a single model. The principal contribution of the paper is to combine recent results on networks of queues of several different service disciplines and a broad class of service time distributions with earlier results on networks of queues containing different classes of customers. We derive the equilibrium state probabilities for the general model. The technique of analysis uses Whittle's concept of independent balance [16,17]. From the complete equilibrium distribution of states of the model, we derive several less complex descriptions of the steady state performance of the model. In the case of certain open networks, we obtain some particularly simple formulas giving the marginal distribution of customers at a service center of the network.

The model is motivated by the conception of a computer system as a network of processors (CPU's, I/O processors, terminals) and a collection of customers (jobs, tasks). The processors are grouped in equivalence classes called service centers and the customers may enter the system from the outside, pass from service center to service center competing for the processing resources of a service center with the other customers at that center, and eventually leave the system. Different service centers may have different scheduling capabilities and different processing resources. Different customers may have different routes through the network and make different demands at a given service center. Customers may change from one class to another when changing service centers. Such a model can represent

several levels of detail in the operation of computer systems, from the job submissions or user logons, through the requests of jobs for individual I/O transfers or computing bursts, to the requests of processors for cycles of a shared memory. We present one example at the middle level of detail.

Several special cases of the model we consider have been studied in the literature. A good survey of the analysis of queueing networks in general and queueing models of computer systems in particular is given by Buzen [3]. Jackson [11] and Gordon and Newell [10] develop the equilibrium distribution of states of a class of general networks. In particular, Gordon and Newell make clear the product form of the solution of the balance equations describing the steady state of the model. Our solution has this product form. In these models the service centers can be connected in any arbitrary fashion. A customer leaving a service center simply chooses the next service center according to a fixed set of branching probabilities for the center being left. Jackson's model also allows for the arrival and departure of customers from outside the system. These networks suffer from two principal limitations as models of computer systems: (1) all the customers are identical; they all follow the same rules of behavior, and (2) all the service time distributions are exponential. These limitations have been attacked by a number of authors. We summarize their results in the remainder of this introduction. The body of the paper presents the general model for which the models discussed below are special cases.

Ferdinand [9] analyzed a particular system which allowed different classes of customers. The system was a cyclic model with two service centers. The model is frequently called the finite source model or the machine repairman model. One service center consists of a sufficient number of servers so that no queueing occurs. The other service center is a single server. There is a fixed number of customers, each of which is characterized by its own pair of exponentially distributed service times, one for each service center. The single server is characterized by processor sharing scheduling in which all waiting customers are processing simultaneously, but at a rate reduced by a factor of 1/n if n customers are requiring service. His solution for the equilibrium distribution of states has the product form. His model is a special case of our model having two service centers, one of a processor sharing type and one of a no queueing type and exponentially distributed service times for the different classes of customers.

Posner and Bernholtz [14] consider the more general network model of Cordon and Newell in which each customer has its own set of branching probabilities, its own set of exponentially distributed service times, and its own generally distributed travel time to a particular service center for each service center in the network. When different customers have different service time distributions at a service center with queueing, processor sharing scheduling is used at that service center. Their model is a special case of our model in that only FCFS and processor sharing types of centers are allowed, the network is closed, and only exponentially distributed service times for the different classes of customers are allowed.

Processor sharing scheduling has been investigated in models of computer systems as the limit of overhead free round-robin scheduling.

The mathematical form of the equations solved by Ferdinand and by Posner and Bernholtz is the form obtained for processor sharing scheduling although neither of the papers clearly identifies the type of processor scheduling being used. Sakata, Noguchi, and Oizumi [15] discovered that when processor sharing scheduling was applied to the classical infinite source queueing model (denoted M/G/1), the equilibrium distribution of queue sizes for the model was the same as that for a similar model with exponentially distributed service times (denoted M/M/1) with the same mean as the original general distribution. Their model is not a special case of the model studied here but can be obtained from it by a limiting argument. Baskett [1] derived a similar result for a finite source model in which the service time distributions at both service centers have rational Laplace transforms and Baskett and Palacios [2] extended that result to another network model which Buzen [3] has studied and called the central server model. The equilibrium solutions have the product form. The models include FCFS, processor sharing, and no queueing types of service centers and service time distributions with rational transforms but only limited closed structure and only a single class of customers.

Whittle [16,17] showed that the balance equations describing interconnected birth and death processes could be replaced by sets of "independent" balance equations and that solutions for these independent sets are solutions for the original equations. Chandy [4,5] showed that this technique could be applied to more complex models and with it he easily extended and generalized earlier results on models with rational service times. Chandy calls these independent sets of equations the equations of local balance and we follow his terminology, The equations of local balance can be written down directly for such models and they are much easier to manipulate and to solve for those models to which they apply. Using this technique, Chandy greatly extended

the range of networks for which product form solutions can be found. In the terminology of this paper, Chandy developed the solution for networks in which the service center is of FCFS, processor sharing, or LCFS type and in which all customers are the same. The results in this paper generalize his results to include service centers of no queueing type and different classes of customers. Palacios [13] independently developed solutions for a particular network with "types" of customers. Chandy, Keller, and Browne [6] then extended the concept of customer "type" and added the concept of customer "mode" for general networks. These concepts can be shown to be equivalent to our classes of customers where customers may change classes.

The next section describes the model and the four types of service centers. Then we discuss distributions with rational **Laplace** transforms. Next is the notation used to indicate the state of the model, a discussion of local balance, and the derivation of the relative frequency with which each class of customers visits each service center. We then give the functional form of the equilibrium state probabilities for the model. This gives a steady state description of the model in more detail than we normally need. The next section develops equilibrium probabilities for composite states of the model. For open models, we obtain a closed form expression for the normalizing constant in the solution and some especially simple formulas for the marginal distribution of customers at each service center. We present numerical results from a closed model with two classes of customers to indicate the significance of different classes of customers. Finally we present an equivalent model in which all classes of customers have the same service time distribution at each service center and all these distributions are exponential.

The Model

The class of systems under consideration contain an arbitrary but finite number N, of service centers. There is an arbitrary but finite number R, of different classes of customers. Customers travel through the network and change class according to transition probabilities. Thus a customer of class r which completes service at service center i will next require service at center j in class s with a certain probability denoted $p_{i,r;j,s}$. Both open and closed networks will be treated. The transition matrix $P = [p_{i,r;j,s}]$ defines a Markov chain where the states are labeled by the pairs (i,r). This Markov chain is assumed to be decomposable into irreducible ergodic subchains. Let $E_1, E_2, \ldots, E_m$ be the sets of states in each of these subchains. Let n_{ir} be the number of customers of class r at service center i. Let $\sum_{(ir)\in E_j} (n_{i,r}) = M(E_j)$. Then in a closed system

$$M(E_j) = \text{constant} \qquad 1 \le j \le m$$

In an open system customers may arrive to the network from an external source. Two general types of state dependent arrival processes are considered. In the first case the total arrival rate to the network is Poisson with mean rate dependent on the total number of customers in the network. Thus for a state S of the model let M(S) be the total number of customers in the network and $\lambda(M(S))$ be the instantaneous mean arrival rate. An arrival enters service station i in class r with a fixed probability (not state dependent) given by q_{ir}.

In the second type of arrival process there are m Poisson arrival streams corresponding to the m subchains defined above. The instantaneous mean arrival rate for the j^{th} stream is assumed to be a function of $M(E_j)$, $\lambda_j(M(E_j))$. An arrival in the j^{th} stream has probability q_{ir} of entering

service station i in class r and $q_{ir} = 0$ if $(i,r) \notin E_j$. In an open network, a customer of type r which completes service at center i may leave the system. This occurs with probability

$$1 - \sum_{\substack{1 \le j \le N \\ 1 \le s \le R}} p_{i,r;j,s}$$

A service center will be referred to as type 1, 2, 3 or 4 according to which condition it satisfies.

Condition 1: There is a single server at a service center, the service discipline is FCFS, all customers have the same service time distribution at this service center, and the service time distribution is a negative exponential with parameter $\mu(n)$, a function of the instantaneous queue size, n, at the server.

Condition 2: There is a single server at a service center, the service discipline is processor sharing (i.e. when there are n customers in the service center each is receiving service at a rate of 1/n sec./sec.), and each class of customer may have a distinct service time distribution. The service time distributions have rational Laplace Transforms.

Condition 3: The number of servers in the service center is greater than or equal to the maximum number of customers that can be queued at this center in a feasible state and each class of customer may have a distinct service time distribution. The service time distributions have rational Laplace Transforms.

Condition 4: There is a single server at a service station, the queueing discipline is preemptive-resume LCFS, and each class of customer may have a distinct service time distribution. The service time distributions have rational Laplace Transforms.

A type one service center with more than one server is equivalent to a type one service center with one server and suitably chosen service rates depending on the number of customers at the server. We denote the service rate at service center i as $\mu_i(j)$ when the center is type one and j customers are awaiting or receiving service at that center.

Representation of Service Time Distributions with Rational **Laplace** Transforms

The requirement that a service time distribution have a rational **Laplace** Transform is not very restrictive. Exponential, hyperexponential and hypoexponential distributions all have rational **Laplace** Transforms. Cox [12] has shown that any such distribution can be represented by a network of exponential stages of the form illustrated in Fig. 1. For convenience, we have eliminated the case in which there is a non-zero probability of a zero length service time.

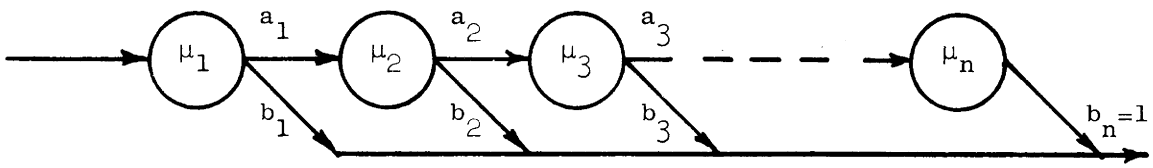


Fig. 1

In this figure, b_i is the probability that the customer leaves after the i^{th} stage and a_i $(=1-b_i)$ is the probability that the customer goes to the next stage. Given that a customer reaches the i^{th} stage the service time in this stage has a negative exponential distribution with mean $1/\mu_i$. Since the service time distribution for a stage is exponential, when describing the state of the network of service stations it is not necessary to know the exact amount of service a customer has received at a service center; the stage of service is sufficient.

The States of the Model

The state of the model is represented by a vector $(x_1, x_2, \ldots, x_N)$ where x_i represents the conditions prevailing at service center i. The interpretation of x_i depends on the type of service center i.

If service center i is of type 1 then

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$$

where n_i is the number of customers at center i and x_{ij} $(1 \leq j \leq n_i, 1 \leq x_{ij} \leq R)$ is the class of customer who is j^{th} in FCFS order. The first customer is served while the remainder are waiting for service.

If service center i is of type 2 or 3 then

$$x_i = (v_{i1}, v_{i2}, \ldots, v_{iR})$$

where v_{ir} is a vector $(m_{1r}, m_{2r}, \ldots, m_{u_{ir}r})$. The ℓ^{th} component of v_{ir} is the number of customers of class r in center i and in the ℓ^{th} stage of service. u_{ir} is the number of stages for a class r customer at service center i.

If service center i is of type 4 then

$$x_i = ((r_1, m_1), (r_2, m_2), \ldots, (r_{n_i}, m_{n_i}))$$

where n_i is the number of customers at center i and (r_j, m_j) is a pair describing the j^{th} customer in LCFS order. r_j is the class of this customer and m_j is the stage of service this customer is in.

For any network of reasonable size, the expression for a state of the network is long and tedious to write. Writing expressions for the balance equations to find the equilibrium state probabilities is an arduous task.

Even to check that a given solution is correct is time consuming. The solution for the class of networks described here was arrived at by using the technique of local balance. This technique is briefly described below.

A solution for the equilibrium state probabilities must satisfy the balance equations for the system. That is

$$\forall \text{ states, } S_i \quad \sum_{\substack{\text{all states} \\ S_j}} P(S_j)[\text{rate of flow from } S_j \text{ to } S_i] = P(S_i)[\text{rate of flow out of } S_i]$$

In [4], Chandy terms these the global balance equations. He defines another type of balance equation which he calls the local balance equations. Informally, a local balance equation equates the rate of flow into a state by a customer entering a stage of service to the flow out of that state due to a customer leaving that stage of service. We associate a customer with a stage of service in the following ways. If the customer is in service at a service center, then he is in one of the stages of his service time distribution at that service center. If the customer is queued at a service center, then he is in the stage of his service time distribution he will enter when next given service. For FCFS this will be stage 1 and for LCFS this will be the stage the customer was in when last preempted.

The local balance equations are sufficient conditions for global balance, but they are not necessary. Local balance requires that each term on the right-hand side of a global balance equation be equal to a particular subset of terms on the left-hand side of that global balance equation.

To illustrate the concept of local balance we consider the relatively simple network model in Fig. 2.

This is a closed network with two classes of customers (which we refer to as class 1 and class 2). There are N_1 class 1 customers and N_2 class 2 customers in the networks. All service times are exponentially distributed and $\frac{1}{\mu_{ir}}$ $(i = 1,2,\ r = 1,2)$ is the mean service time for a class r customer at service center i.

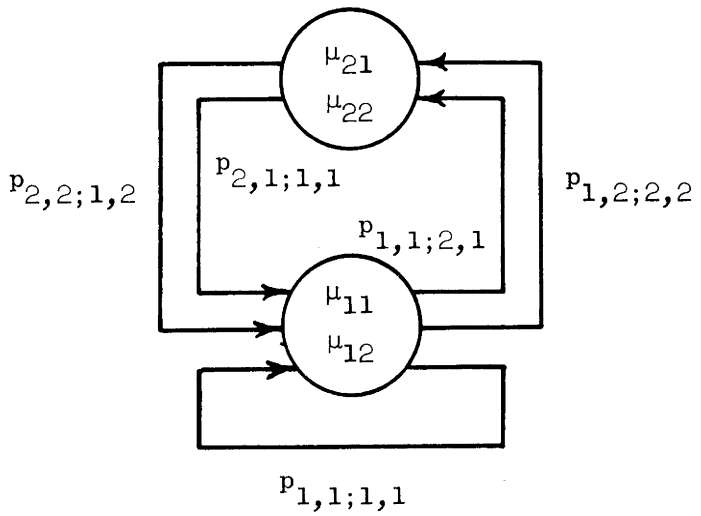


Fig. 2

In this example $p_{1,_{2:2}',_2} = p_{2,_{2:1}',_2} = p_{2,1;1,1} = 1$, $p_{1,1;1,1} + p_{1,1;2,1} = 1$.

Let $n_{i,r}$ be the number of class r customers at service center i. For convenience we write the global and local balance equations only for the states in which $n_{ir} > 0$, i = 1,2, r = 1,2.

Global Balance Equation:

$$P(n_{11}-1,n_{12},n_{21}+1,n_{22})\left(\frac{n_{21}+1}{n_{21}+n_{22}+1}\right)\mu_{21}$$

$$+ P(n_{11}+1,n_{12},n_{21}-1,n_{22})\,(n_{11}+1)\,\mu_{11}\,p_{1,1;2,1}$$

$$+ P(n_{11},n_{12},n_{21},n_{22})\,n_{11}\,\mu_{11}\,p_{1,1;1,1}$$

$$+ P(n_{11},n_{12}+1,n_{21},n_{22}-1)\,(n_{12}+1)\,\mu_{12}$$

$$+ P(n_{11},n_{12}-1,n_{21},n_{22}+1)\left(\frac{n_{22}+1}{n_{21}+n_{22}+1}\right)\mu_{22}$$

$$= P(n_{11},n_{12},n_{21},n_{22})\left[n_{11}\,\mu_{11} + n_{12}\,\mu_{12} + \frac{n_{21}}{n_{21}+n_{22}}\,\mu_{21} + \frac{n_{22}}{n_{21}+n_{22}}\,\mu_{22}\right]$$

Local Balance Equations:

(1.1) $$P(n_{11}-1,n_{12},n_{21}+1,n_{22})\left(\frac{n_{21}+1}{n_{21}+n_{22}+1}\right)\mu_{21} + P(n_{11},n_{12},n_{21},n_{22})\,n_{11}\,\mu_{11}\,p_{1,1;1,1} = P(n_{11},n_{12},n_{21},n_{22})n_{11}\,\mu_{11}$$

(1.2) $$P(n_{11},n_{12}-1,n_{21},n_{22}+1)\left(\frac{n_{22}+1}{n_{21}+n_{22}+1}\right)\mu_{22} = P(n_{11},n_{12},n_{21},n_{22})n_{12}\,\mu_{12}$$

(2.1) $$P(n_{11}+1,n_{12},n_{21}-1,n_{22})(n_{11}+1)\mu_{11}\,p_{1,1;2,1} = P(n_{11},n_{12},n_{21},n_{22})\left(\frac{n_{21}}{n_{21}+n_{22}}\right)\mu_{21}$$

(2.2) $$P(n_{11},n_{12}+1,n_{21},n_{22}-1)(n_{12}+1)\mu_{12} = P(n_{11},n_{12},n_{21},n_{22})\left(\frac{n_{22}}{n_{21}+n_{22}}\right)\mu_{22}$$

Since all the service time distributions in this example are exponential the current stage of service of a customer is uniquely defined by the customer's class and the current service center. Local balance equation (i.r) for $i = 1,2$, $r = 1,2$ equates the rate of flow out of state $(n_{11},n_{12},n_{21},n_{22})$ due to a class r customer leaving service center i with the rate of flow into state $(n_{11}, n_{12},n_{21},n_{22})$ due to a class r customer entering service center i.

As in this example it is generally true that each global balance equation is the sum of a subset of the local balance equations. Thus a solution for the local balance equations is automatically a solution to the global balance equations. In many cases the local balance equations are inconsistent and therefore have no solution. For example if there is FCFS

scheduling at a service center and different classes of customer have different service time distributions the local balance equations are inconsistent.

The value of the local balance concept is that (1) it leads to a simpler and more organized search for solutions for equilibrium state probabilities and (2) it works for a large number of cases (in fact for virtually all of the closed form solutions known for general classes of networks of queues--although not many interesting cases have known solutions).

Before presenting the solution to the class of networks described, we define a set of terms that appear in the solution.

For a customer of class r, let $\{e_{ir}, 1 \leq i \leq N)$ be a solution to the following set of equations.

$$\sum_{1 \leq i \leq N} e_{ir} p_{i,r;j,s} + d_{Js} = e_{Js} \quad , \qquad 1 \leq j \leq N$$

The value of d_{js} is determined by the arrival process of customers of class s to service center j. If there are no such arrivals from outside the system, then $d_{Js} = 0$. If there are such arrivals then $d_{js} = q_{js}$. In a closed system there are no arrivals to any center and all the d_{Js} are zero. In this case e_{ir} is the relative frequency of visits to service center i by customers of class r.

Note that a system may be "open' w.th respect to some classes of customers and "closed" with respect to other classes of customers. Our solution applies to this class of system.

One further definition is required. If at the i^{th} service center the r^{th} class of customers has a service time distribution that is represented as a network of stages then this is represented as illustrated in Figure 3.

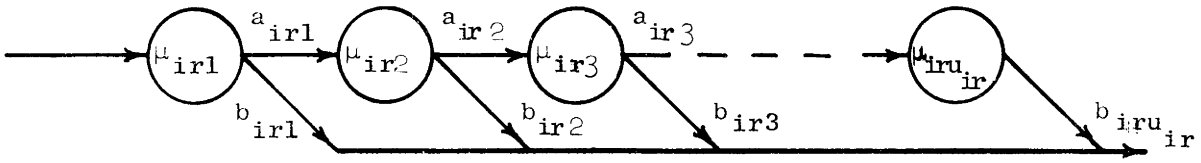


Fig. 3

The first subscript on a, b and μ denotes the service center. The second subscript denotes the class of customer and the third subscript denotes the stage.

Let $A_{ir\ell} = \prod_{j=1}^{\ell} a_{irj}$

Theorem:

Given a network of service stations which is open, closed or mixed in which each service center is of type 1, 2, 3 or 4. Then the equilibrium state probabilities are given by

$$P(S = x_1, x_2, \ldots, x_N) = Cd(S)\, f_1(x_1) f_2(x_2) \ldots f_N(x_N)$$

where C is a normalizing constant chosen to make the equilibrium state probabilities sum to 1, d(S) is a function of the total number of customers in system and each f_i is a function that depends on the type of service center i.

If service center i is of type 1 then

$$f_i(x_i) = \prod_{j=1}^{n_i} \left[\frac{1}{\mu_i(j)}\, e_{ix_{ij}} \right]$$

If service center i is of type 2 then

$$f_i(x_i) = n_i! \prod_{r=1}^{R} \prod_{\ell=1}^{u_{ir}} \left\{ \left[\frac{e_{ir} A_{ir\ell}}{\mu_{ir\ell}} \right]^{m_{ir\ell}} \frac{1}{m_{ik\ell}!} \right\}$$

If service center i is of type 3 then

$$f_i(x_i) = \prod_{r=1}^{R} \prod_{\ell=1}^{u_{ir}} \left\{ \left[\frac{e_{ir} A_{ir\ell}}{\mu_{ir\ell}} \right]^{m_{ir\ell}} \frac{1}{m_{ir\ell}!} \right\}$$

If service center i is of type 4 then

$$f_i(x_i) = \prod_{j=1}^{n_i} \left[e_{ir_j} A_{ir_j} \frac{1}{\mu_{ir_j} m_j} \right]$$

If the arrivals to the system depend on the total number of customers in the system, M(S), and the arrivals are of class **r** and for center i according to fixed probabilities p_{ir} then

$$d(S) = \prod_{i=0}^{M(S)-1} \lambda(i)$$

If we have the second type of state dependent arrival process then

$$d(S) = \prod_{j=1}^{m} \prod_{i=0}^{M(E_j)-1} \lambda_j(i)$$

If the network is closed then d(S) = 1.

The theorem is proved by checking that the local balance equations are satisfied. In every case for which these results apply the local balance equations reduce to the defining equations for the $\{e_{ir}\}$.

Simplification of Results

The solution presented for the equilibrium state probabilities deals with system states that are more detailed than is usually required. The more detailed states are necessary to derive the equilibrium state probabilities. Now we define the system state as the number of each class of customer in each service center. More formally state S of the system is given by $(y_1, y_2, \ldots, y_N)$ where $y_i = (n_{i1}, n_{i2}, \ldots, n_{iR})$ and nir is the number of customers of class r in service center i. Let n_i be the total number of customers at service center i and let $\frac{1}{\mu_{ir}}$ be the mean service time of a class r customer at service center i. Then the equilibrium state probabilities are given by

$$P(S = (y_1, y_2, \ldots, y_N)) = Cd(S)\ g_1(y_1) g_2(y_2) \ldots g_N(y_N)$$

where

if service center i is of type 1 then

$$g_i(y_i) = n_i! \left\{ \prod_{r=1}^{R} \frac{1}{n_{ir}!} [e_{ir}]^{n_{ir}} \right\} \prod_{j=1}^{n_i} \frac{1}{\mu_i(j)}$$

if service center i is of type 2 or 4 then

$$g_i(y_i) = n_i! \prod_{r=1}^{R} \frac{1}{n_{ir}!} \left[\frac{e_{ir}}{\mu_{ir}} \right]^{n_{ir}}$$

if service center i is of type 3 then

$$g_i(y_i) = \prod_{r=1}^{R} \frac{1}{n_{ir}!} \left[\frac{e_{ir}}{\mu_{ir}} \right]^{n_{ir}}$$

In each case the expression for $g_i(y_i)$ is derived by summing $f_i(x_i)$ over all x_i with $n_{i1}, n_{i2}, \ldots, n_{ik}$ fixed. That this is the correct definition of the g_i follows from the product form of the solution given

in the theorem. If the mean service rate at centers of type 2, 3, or 4 is the same for each class of jobs but depends on the number of customers at the center, then the factor $\prod_{r=1}^{R} \left(\frac{1}{\mu_{ir}}\right)^{n_{ir}}$ is replaced by the product of the $\frac{1}{\mu_i(j)}$ in $g_i(y_i)$ as for type one centers. If the service rates are not the same for each class of customers but depend on the number of customers at a center, then it should be possible to develop the proper form of the solution using the equivalent network presented in the last section of the paper.

A further simplification is possible if the network is open and the arrival process does not depend on the state of the model, The following paragraph and section develop this simplification,

If a state of the system is to be simply the total number of customers in each service station, i.e. $S = (n_1, n_2, \ldots, n_N)$. Then $P(S) = Cd(S)\ h_1(n_1)\ h_2(n_2)\ldots h_N(n_N)$. Let R_i = (r: class r customers may require service center i).

If service station i is of type 1 then

$$h_i(n_i) = \left(\sum_{r \in R_i} e_{ir}\right)^{n_i} \prod_{j=1}^{n_i} \frac{1}{\mu_i(j)}$$

If service station i is of type 2 or 4 then

$$h_i(n_i) = \left(\sum_{r \in R_i} \frac{e_{ir}}{\mu_{ir}}\right)^{n_i}$$

If service station i is of type 3 then

$$h_i(n_i) = \frac{1}{n_i!}\left(\sum_{r \in R_i} \frac{e_{ir}}{\mu_{ir}}\right)^{n_i}$$

The evaluation of the normalizing constant requires summing the given expression for the equilibrium state probabilities over all feasible states. In the next section we show a closed form solution for C for an open network in which $\mu_i(m) = \mu_i$ for all m if service center i is of type one.

Open Systems

For open systems it is possible to obtain a closed form solution for the normalization constant when the arrival process is of the first type and $\lambda(M(S)) = \lambda$ = constant. Since the system is open any number of customers is feasible at a service center.

Therefore

$$C^{-1} = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \cdots \sum_{n_N=0}^{\infty} \left(\prod_{i=1}^{N} \lambda^{n_i} h_i(n_i) \right)$$

$$\text{or } C^{-1} = \left(\sum_{n_1=0}^{\infty} \lambda^{n_1} h_1(n_1) \right) \left(\sum_{n_2=0}^{\infty} \lambda^{n_2} h_2(n_2) \right) \cdots \left(\sum_{n_N=0}^{\infty} \lambda^{n_N} h_N(n_N) \right)$$

Also,

$$\sum_{n_i=0}^{\infty} h_i(n_i) = \left(1 - \sum_{r \in R_i} \lambda \frac{e_{ir}}{\mu_i} \right)^{-1} \quad \text{if service center i is type 1 and } \mu_i(n_i) = \mu_i$$

$$= \left(1 - \sum_{r \in R_i} \lambda \frac{e_{ir}}{\mu_{ir}} \right)^{-1} \quad \text{if service center i is type 2 or 4}$$

$$= e^{\sum_{r \in R_i} \lambda \frac{e_{ir}}{\mu_{ir}}} \quad \text{if service center i is type 3}$$

Marginal Distribution at a Service Center in an Open System

Let $P_i(n_i)$ be the equilibrium probability that there are n_i customers at service center i.

$$P_i(n_i) = C\lambda^{n_i} h_i(n_i) \prod_{\substack{j=1 \\ j\neq i}}^{N} \left(\sum_{n_j=0}^{\infty} \lambda^{n_j} h_j(n_j) \right)$$

Using the expression for C, we reduce this to

$$P_i(n_i) = \frac{\lambda^{n_i} h_i(n_i)}{\sum_{m=0}^{\infty} \lambda^m h_i(m)}$$

Let $\rho_i = \sum_{r\in R_i} \lambda \frac{e_{ir}}{\mu_i}$ if service center i is type 1

$\rho_i = \sum_{r\in R_i} \lambda \frac{e_{ir}}{\mu_{ir}}$ if service center i is type 2, 3 or 4

Then $P_i(n_i) = (1 - \rho_i)\rho_i^{n_i}$ if service center i is type 1, 2 or 4

$= e^{-\rho_i} \frac{P_i^{n_i}}{n_i!}$ if service center i is type 3

These results provide a convenient way of examining the equilibrium distribution at a service center. For type 1, 2 or 4 service stations the marginal distribution is the same as the distribution of the number of customers in an M/M/1 queue with a suitably chosen utilization, ρ_i. For the equilibrium solution to exist each ρ_i is required to be less than 1.

The marginal distribution for a type 3 service center is the same as the equilibrium distribution for the number of customers for an M/G/∞ system with $\rho_i = \frac{\lambda}{\mu}$. This certainly appears to be reasonable since for an open

system there must be an infinite number of servers at center i if it is to be of type 3.

This type of service station may be used in a model to represent a delay as customers travel between two other service centers. Posner and Bernholtz [14] use a different approach to represent more general delays in a less general network.

Example

In this section we give a simple example that illustrates some of the results of the paper. Consider the system shown in Figure 4.

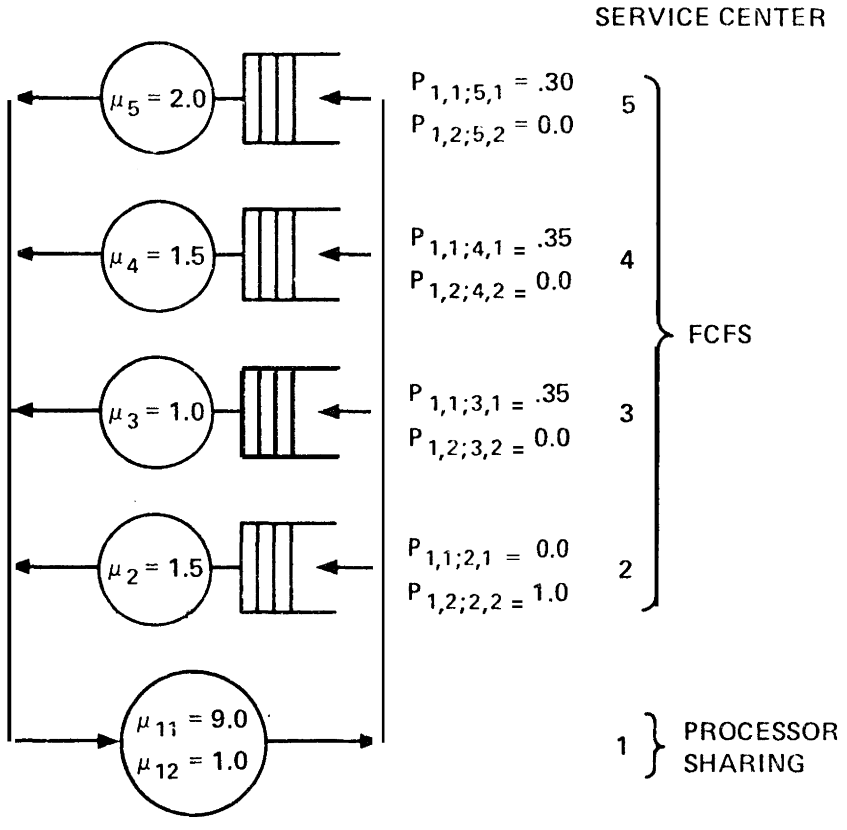


Figure 4. Example Network Model.

This is a **closed** system with two classes of customers. Service centers 2, 3, 4 and 5 are type 1 centers and service center 1 is type 2. This is a model of a multiprogrammed computer system in which service center 1 represents the CPU and the other service centers represent I/O devices.

Figure 5(a) gives the utilizations of the service centers with a varying number of class 1 customers and with one class 2 customer in the system. In Figure 5(b) the utilizations of the service centers are given for the same network of service centers but with the two classes of customers replaced

by one class of "equivalent?' customers. The parameters for these "equivalent" customers are calculated by first solving for the equilibrium state probabilities of the two customer class model. From these one can solve for

r_1 = rate at which class 1 customers leave service center 1.

r_2 = rate at which class 2 customers leave service center 1.

Now the equivalent customers have parameters given by

$$\frac{1}{\mu_1} = \frac{r_1}{r_1+r_2}\ \frac{1}{\mu_{11}} + \frac{r_2}{r_1+r_2}\ \frac{1}{\mu\mu_{12}}$$

$$p_{1;i} = \frac{r_1}{r_1+r_2}\ p_{1,1;i,1} + \frac{r_2}{r_1+r_2}\ p_{1,2;i,2} \qquad i=2,3,4,5$$

The rationale for these definitions is quite simple. If measurements were taken on the system without distinguishing between classes of customers these would be the parameters measured.

Figure 6 shows the results of Fig. 5 graphically. The service center utilizations for the model with different customers are indicated by a line through the values with the service center number above the line. For the model with "equivalent" customers, the service center number is primed and below the line. The utilizations predicted by the model with equivalent customers are always smaller than those of the model with distinct customers. In fact the utilization of service center one (the CPU) goes down initially as the number of "equivalent" customers increases from one to two and the difference for this server is substantial (between 4.5 and 9 percent). The structure of the model with different customers is such that the class 2 customer never has to queue for any I/O server. In the model with equivalent customers, all customers suffer queueing delays at I/O servers for two or more customers.

	UTILIZATIONS OF SERVICE CENTERS				
	1	2	3	4	5
N1 = 0	.600	.400	0.0	0.0	0.0
N1 = 1	.678	.371	.384	.256	.165
N1 = 2	.720	.352	.606	.404	.260
N1 = 3	.744	.339	.743	.495	.318
N1 = 4	.759	.330	.831	.554	.356
N1 = 5	.769	.324	.888	.592	.381
N1 = 6	.775	.321	.926	.617	.397
N1 = 7	.779	.318	.951	.634	.407

(a)

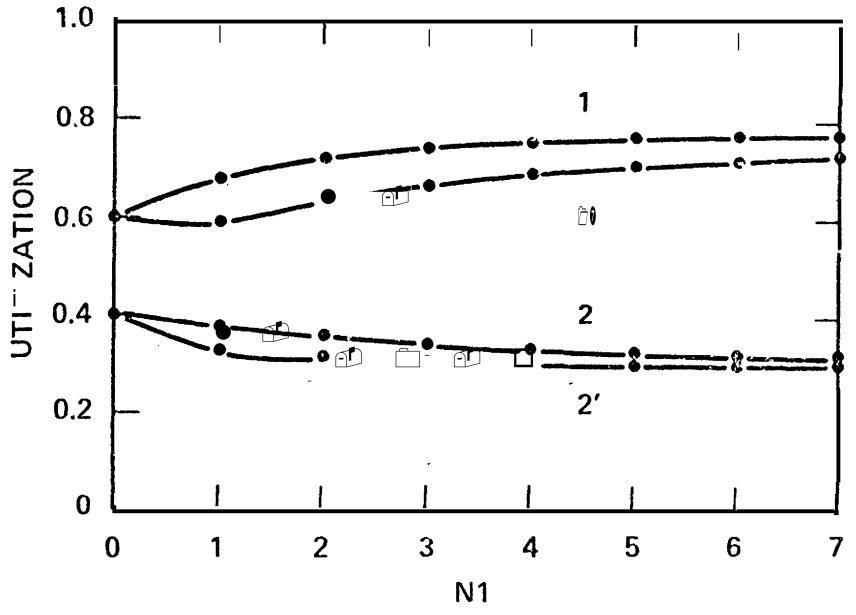
TWO CLASSES OF CUSTOMERS
N2 = # OF CLASS 2 CUSTOMERS
= 1

	UTILIZATIONS OF SERVICE CENTERS					TRANSITION PROBABILITIES				
	1	2	3	4	5	P2/1	P3/1	P4/1	P5/1	μ_1
N1 = 0	.600	.400	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
N1 = 1	.588	.322	.333	.222	.143	.336	.232	.232	.199	2.439
N1 = 2	.631	.308	.532	.354	.228	.233	.268	.268	.230	3.139
N1 = 3	.665	.303	.664	.442	.284	.193	.282	.282	.242	3.536
N1 = 4	.689	.300	.754	.503	.323	.173	.290	.290	.248	3.780
N1 = 5	.708	.299	.818	.545	.350	.161	.294	.294	.252	3.934
N1 = 6	.722	.299	.863	.575	.370	.154	.296	.296	.254	4.034
N1 = 7	.734	.300	.896	.597	.384	.149	.298	.298	.255	4.100

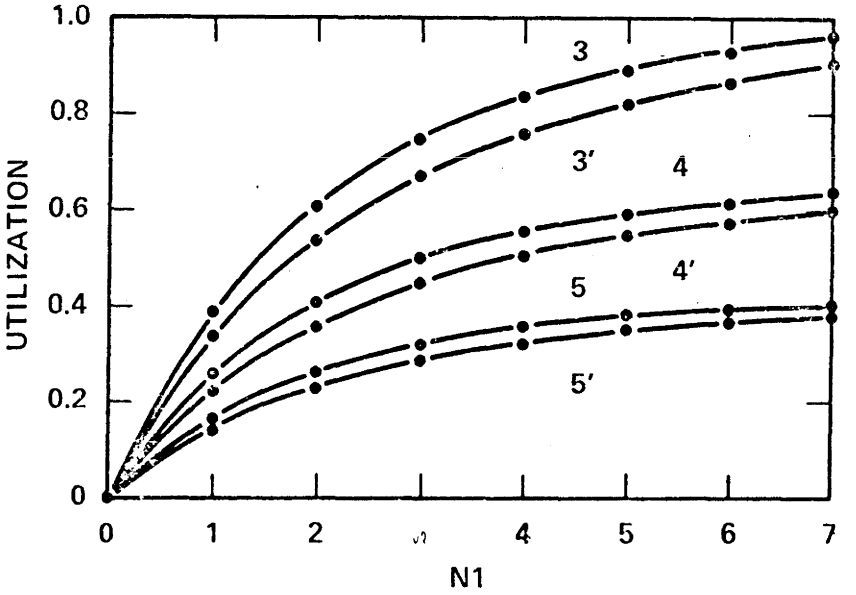
(b)

SAME SYSTEM WITH ONE CLASS OF
"EQUIVALENT" CUSTOMERS
NO. OF CUSTOMER = N1 + N2

Fig. 5



(a)



(b)

Figure 6(a) and (b). Utilization of Service Centers versus Number of Customers for Different Customers and Equivalent Customers.

The customer class change concept can also be used to capture some complex structural properties of the system being modeled. For example, in one of Moore's [12] models of a timesharing system one drum service time is used in the model to represent two drum service times in the system. One of the service times is incurred by the transfer of a job from terminal I/O processing to CPU and file processing and the other corresponds to the reverse transfer, A more accurate representation of contention on a swapping drum can be obtained by using two classes for jobs. One class would model terminal I/O processing and the other would model CPU and file processing. A job would change class after drum processing.

Properties of Network Models that Satisfy Local Balance

This section is directed to those readers interested in the theoretical foundations of the analysis of networks of queues rather than those interested in the application of the results.

All of the network models that we have treated in this paper can be shown to be equivalent to models in which all classes of customers have the same exponential service time distribution at a given service center. Thus an exponential service time distribution with mean $\frac{1}{\mu_i}$ may be associated with the i^{th} service center and all classes of customers have this service time distribution at the i^{th} service center. This fact suggests the conjecture that a necessary condition for local balance to be satisfied for a given model is that there exist an equivalent model in which different classes of customers may have different transition probabilities but all classes of customers have the same exponential service time distribution at a given service center.

The transformation of a given model to an equivalent model of the form described is accomplished in two steps. First we show that the effect of a customer moving from one stage to another in the stages representation of a general service time distribution can be represented by introducing new customer classes. Thus we model a transition from one stage to the next as a transition to a new customer class and to the same service center. After this transformation of the original model we have a model in which all service times are exponentially distributed but different classes of customers may have different mean service times at a given service center. The second step is to show that by appropriately modifying transition probabilities we may further transform the model into an equivalent model

in which all classes of customers have the same mean service time at a given service center.

The method of making these transformations to the model is straightforward and will be illustrated by example rather than a formal description of the general case.

Consider a general service time distribution represented by a network of stages as in Figure 1. Let this represent the service time distribution for a customer in class r in service center i. We introduce n new customer classes denoted by $r_1, r_2, \ldots, r_n$ which correspond to the stages in this network and delete customer class r.

The service time of a class r_ℓ customer will be exponential with mean $\frac{1}{\mu_\ell}$ $(1 \leq \ell \leq n)$. The transition probabilities for a class r_ℓ customer are defined as:

$$p_{i,r_\ell;j,s} = \beta_\ell \; p_{i,r;j,s}$$

$$p_{i,r_\ell;i,r_{\ell+1}} = a_\ell \qquad 1 \leq \ell < n$$

To take care of the transitions into class r in the original model we require that all transitions into state r be redefined as transitions into state r_1. These transition probabilities are defined as:

$$p_{j,s;i,r_1} = p_{j,s;i,r} \;, \quad \forall j,s$$

With this transformation of the model a customer will have the same distribution of total time at a service center and will have the same transition probabilities from service center to service center.

After performing this transformation for each customer class with a general service time distribution we have a model in which all service time distributions are exponential. Suppose that $\frac{1}{\mu_{i,r}}$ is the mean service time of a class **r** customer at service center i. Let

$$\mu_i = \max_r \{M_{i,t}\}$$

We redefine the mean service time for each class of customers at service center i to be $\frac{1}{\mu_i}$. Now we redefine the transition probabilities out of service center i.

Let $p_r = 1 - \frac{\mu_{i,r}}{\mu_i}$.

Then define

$$P'_{i,r;i,r} = p_r + (1-p_r)\, P_{i,r;i,r}$$

$$P'_{i,r;j,s} = (1-p_r) p_{i,r;j,s}$$

The effect of these new transition probabilities is to cause a class **r** customer to be fed back (or to revisit) service center i a random number of times. Each time the class **r** customer enters service center i his service time is exponentially distributed with mean $\frac{1}{\mu_i}$. The number of visits the class **r** customer makes to service center i (between transitions in the original model) is geometrically distributed with mean $\frac{1}{1-p_r}$. It is easily shown that the total service time of the class **r** customer at service center i is exponentially distributed with mean $(\frac{1}{1-p_r})\frac{1}{\mu_i} = \frac{1}{\mu_{i,r}}$ [8]. Therefore we have not changed the total service time distribution for this customer at service center i.

After completing these transformations throughout the model we have an equivalent model with the desired characteristics.

The transformations that we have made to the original model preserved the original distributions of service time that a customer requires at a service center. However a customer does not spend that time on the server in one contiguous interval. We required a customer to make extra transitions in which he leaves and reenters the service center. It is clear that with type 2, 3 or 4 service centers this does not affect a customers service. For type 1 service centers the transformed model would not be equivalent to the original model since a customer who leaves the service center and reenters will now be at the end of the queue. Of course we have from the beginning required that at type 1 service centers all customers have the same exponential service time distribution so that such a service center does not require any modification.

Conclusions

We have derived the equilibrium distribution of states of a model containing four different types of service centers and R different classes of customers. From this steady state distribution one can compute the moments of the queue sizes for different classes of customers at different service centers, the utilizations of the service centers, the "cycle time" or response time for different classes of customers, the "throughput" of different classes of customers, and other measures of system performance.

These results unify and extend a number of separate results on networks of queues. The general model can have four types of service centers. Three of those types allow different service time distributions with rational **Laplace** transforms for different classes of customers. The model allows different classes of customers to have different arrival rates and different routing probabilities. For open networks some very simple formulas give the marginal distribution of customers at the service centers of the network.

The analysis is motivated by the desire to model computer systems. Type one service centers (FCFS scheduling) seem appropriate models of secondary storage input/output devices. Type two service centers (processor sharing scheduling) can be an appropriate model for central processing units. Type three service centers (no queueing) are appropriate models for terminals and for routing delays in the network. Allowing different classes of customers should answer one of the principal objections to queueing models as models of computer systems. The example given indicates how significant different classes of customers can be in the utilization levels predicted by model analysis.

There are many additional complications yet to be analyzed but the general model presented here represents a substantial increase in the ability to build and solve analytical models of complex computer systems.

Acknowledgements

This work was supported by Joint Services Electronics Programs: U.S. Army, U.S. Navy, U.S. Air Force under contract N-00014-67-A-0112-0044 and by the Advanced Research Projects Agency of the Department of Defense under contract DAHC-13-69-C-0238,

REFERENCES

1. Baskett, F. The dependence of computer system queues upon processing time distribution and central processor scheduling, Proc. of the ACM SIGOPS Third Symposium on Operating System Principles, Stanford University, (October 1971), 109-113.

2. Baskett, F. and Palacios, F. G. Processor sharing in a central server queueing model of multiprogramming with applications, Proc. of the Sixth Annual Princeton Conference on Information Sciences and Svstems. Princeton University, (March 1972).

3. Buzen, J. Queueing Network Models of Multiprogramming, Ph.D. Thesis, Division of Engineering and Applied Science, Harvard University, Cambridge, Mass., 1971.

4. Chandy, K. M. The analysis and solutions for general queueing networks, Proc. of the Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton University, (March 1972).

5. Chandy, K. M. Exponential and processor-sharing queueing network models for computer systems, IEEE Trans. on Computers, (to appear).

6. Chandy, K. M., Keller, T. W., and Browne, J. C. Design automation and queueing networks: an interactive system for the evaluation of computer queueing models, Proc. of the Ninth Annual Design Automation Workshop, Dallas, June, 1972.

7. Cox, D. R. A use of complex probabilities in the theory of stochastic processes, Proc. Camb. Phil, Sox 51 (1955), 313-319.

8. Feller, W. An Introduction to Probability Theory and its Applications, Vol. 1, 3rd ed., Wiley, New York, 1968.

Ferdinand. A. E. An analysis of the machine interference model, IBM Systems J., 10, 2 (1971), pp. 129-142.

1 . Gordon, W. J. and Newell, G, F. Closed queueing systems with exponential servers, Opns. Res,, 15 (1967), 254-265.

11 Jackson, J. R. Jobshop-like queueing systems, Man. Sci. 10, 1 (Oct. 1963), pp. 131-142.

12 Moore, C. G., III. Network Models for Large-Scale Time-Sharing Systems, Technical Report No. 71-1, Department of Industrial Engineering, The University of Michigan, Ann Arbor, Michigan, (April 1971).

13 Palacios, F. G. An analytic model of a multiprogramming system including a job mix, Report TR-4, Department of Computer Sciences, University of Texas at Austin, June, 1972.

14. Posner, M. and Bernholtz, B. Closed finite queueing networks with time lags and with several classes of units, Opns. Res. 16 (1968), pp. 977-985.

15 Sakata, M., Noguchi, S., and Oizumi, J., Analysis of a processor-shared queueing model for time-sharing systems, Proc. of the Second Hawaii International Conference on System Sciences, January 1969.

16. Whittle, P. Nonlinear migration processes. Proc. 36th Session of the International Statistical Institute, pp. 642-647.

17. Whittle, P. Equilibrium distributions for an open migration process. J. Appl. Prob. 5, pp. 567-571.