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

FOREST BASKETT

Stanford University, Stanford, California

K. MANI CHANDY

University of Texas at Austin, Austin, Texas

RICHARD R. MUNTZ

University of California, Los Angeles, California

AND

FERNANDO G. PALACIOS

University of Texas at Austin, Austin, Texas

ABSTRACT The joint equilibrium distribution of queue sizes in a network of queues containing N service centers and R classes of customers is derived The equilibrium state probabilities have the general form $P(S) = Cd(S)\, f_1(x_1)f_2(x_2) \cdot f_N(x_N)$, 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 It is assumed that the equilibrium probabilities exist and are unique Four types of service centers to model central processors, data channels, terminals, and routing delays are considered 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, state dependent arrival processes are considered Closed networks are those with no exogenous 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. Examples show how different classes of customers can affect models of computer systems.

KEY WORDS AND PHRASES: networks of queues, theory of queues, queueing theory, multiprogramming, time-sharing, processor sharing, Markov processes

CR CATEGORIES: 4 32, 5.5, 6 20

This research was supported in part by the U S. Army, U S. Navy, and U S Air Force Joint Services Electronics Programs under Contract N-00013-67-A-0112-0044, in part by the National Science Foundation under Grants GJ-1084 and GJ-35109, and in part by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC-15-69-C-0158

Authors' addresses. F Baskett, Departments of Computer Science and Electrical Engineering, Stanford University, Stanford, CA 94305; R R Muntz, Department of Computer Science, University of California, Los Angeles, CA 90024, K M Chandy and F G Palacios, Department of Computer Sciences, University of Texas at Austin, Austin, TX 78712

1. *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 [17, 18]. 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 (CPUs, I/O processors, terminals) and a collection of customers (jobs, tasks). The processors are grouped into 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 two examples 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.

Special cases of the results presented here have been developed by Ferdinand [9], Posner and Bernholtz [15], Baskett [1], Baskett and Palacios [2], and Chandy et al. [6]. Sakata et al. [16] developed a related result on processor sharing. Whittle [17, 18] describes the "independent balance equations" technique that simplifies the problem of finding steady state solutions for these networks. Chandy [5] also describes this technique and calls it the principle of local balance.

Section 2 describes the model and the four types of service centers, distributions with rational Laplace transforms, and the notation used to indicate the state of the model. Section 3 is a discussion of independent balance, the derivation of the relative frequency with which each class of customers visits each service center, and 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. Section 4 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. Section 5 discusses state dependent service rates. Section 6 presents two examples to indicate the significance of different classes of customers.

2. *The Model*

2.1. SERVICE CENTERS. The class of systems under consideration contains 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 who completes service at service center i will next require service at center j in class s with a certain probability denoted $P_{i,r,j,s}$. The transition matrix $P = [P_{i,r,j,s}]$ can be considered as defining a Markov chain whose states are labeled by the pairs (i, r). The Markov chain is assumed to be decomposable into m ergodic subchains. Let $E_1, E_2, \cdots, E_m$ be the sets of states in each of these subchains. The possible states of a network model are described in Section 2.3. Let n_{ir} be the number of customers of class r at service center i in state S of the network model. Let $M(S/E_j) = \sum_{(i,r)\in E_j} n_{ir}$. Then a closed system is characterized by $M(S/E_j) = \text{constant}$, $1 \leq j \leq m$.

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 network model let $M(S)$ be the total number of customers in the network, i.e. $M(S) = \sum_{j=1}^{m} M(S/E_j)$, and let $\lambda(M(S))$ 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 jth stream is assumed to be a function of $M(S/E_j)$, $\lambda_j(M(S/E_j))$. An arrival in the jth stream has probability q_{ir} of entering service station i in class r if $(i, r) \in E_j$ and $\sum_{(i,r)\in E_j} q_{ir} = 1$. In an open network a customer of class r who completes service at center i may leave the system. This occurs with probability $1 - \sum_{1\leq j\leq N,\ 1\leq s\leq R} p_{i,r,j,s}$.

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

Condition 1. The service discipline is first-come-first-served (FCFS); all customers have the same service time distribution at this service center, and the service time distribution is a negative exponential. The service rate can be state dependent where $\mu(j)$ will denote the service rate with j customers at the center.

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 center, the queuing discipline is preemptive-resume last-come-first-served (LCFS), and each class of customer may have a distinct service time distribution. The service time distributions have rational Laplace transforms.

Note. An exponential FCFS *single job class* service center with more than one server

is equivalent to a similar service center with one server and suitably chosen service rates depending on the number of customers at the server.

2.2. 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 [7] has shown that any such distribution can be represented by a network of exponential stages of the form shown in Figure 1. For convenience, we have eliminated the case in which there is a nonzero probability of a zero length service time.

In Figure 1, b_i is the probability that the customer leaves after the ith stage and a_i $(=1 - b_i)$ is the probability that the customer goes to the next stage. Given that a customer reaches the ith 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.

2.3. The States of the Model. The state of the model is represented by a vector $(x_1, x_2, \cdots, x_N)$ 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 $x_i = (x_{i1}, x_{i2}, \cdots, x_{in_i})$, 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 jth 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 $x_i = (v_{i1}, v_{i2}, \cdots, v_{iR})$, where v_{ir} is a vector $(m_{1r}, m_{2r}, \cdots, m_{u_{ir}r})$. The lth component of v_{ir} is the number of customers of class r in center i and in the lth 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 $x_i = ((r_1, m_1), (r_2, m_2), \cdots, (r_{n_i}, m_{n_i}))$, where n_i is the number of customers at center i and (r_j, m_j) is a pair describing the jth 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 independent balance. This technique is briefly described in Section 3.

3. *The Equilibrium State Probabilities*

3.1. The Balance Equations. A solution for the equilibrium state probabilities must satisfy the balance equations for the system. That is,

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

Chandy [5] terms these the global balance equations. Whittle [17, 18] describes another type of balance equations which he calls the *independent balance equations*. Informally, an

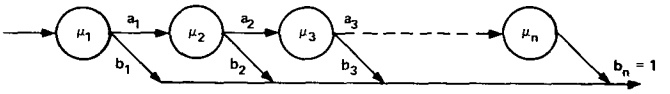

Fig. 1. Representation of service time distributions by the method of stages

independent 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.

From this description of the independent balance equations it is easily seen that each global balance equation is a sum of independent balance equations. Therefore, the independent balance equations are sufficient conditions for global balance (but they are not necessary).

To illustrate the technique of independent balance we consider the relatively simple network model shown in Figure 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 $1/\mu_{ir}$ ($i = 1, 2$, $r = 1, 2$) is the mean service time for a class r customer at service center i.

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

Global Balance Equation:

$$\begin{aligned}
&P(n_{11}-1, n_{12}, n_{21}+1, n_{22})((n_{21}+1)/(n_{21}+n_{22}+1))\mu_{21} \\
&\quad + P(n_{11}+1, n_{12}, n_{21}-1, n_{22})(n_{11}+1)\mu_{11}p_{1,1,2,1} \\
&\quad + P(n_{11}, n_{12}, n_{21}, n_{22})n_{11}\mu_{11}p_{1,1,1,1} \\
&\quad + P(n_{11}, n_{12}+1, n_{21}, n_{22}-1)(n_{12}+1)\mu_{12} \\
&\quad + P(n_{11}, n_{12}-1, n_{21}, n_{22}+1)((n_{22}+1)/(n_{21}+n_{22}+1))\mu_{22} \\
&= P(n_{11}, n_{12}, n_{21}, n_{22})[n_{11}\mu_{11} + n_{12}\mu_{12} + (n_{21}/(n_{21}+n_{22}))\mu_{21} + (n_{22}/(n_{21}+n_{22}))\mu_{22}].
\end{aligned}$$

Independent Balance Equations:

$$\begin{aligned}
&P(n_{11}-1, n_{12}, n_{21}+1, n_{22})((n_{21}+1)/(n_{21}+n_{22}+1))\mu_{21} \\
&\quad + P(n_{11}, n_{12}, n_{21}, n_{22})n_{11}\mu_{11}p_{1,1,1,1} \\
&= P(n_{11}, n_{12}, n_{21}, n_{22})n_{11}\mu_{11} \qquad (1.1)
\end{aligned}$$

$$\begin{aligned}
&P(n_{11}, n_{12}-1, n_{21}, n_{22}+1)((n_{22}+1)/(n_{21}+n_{22}+1))\mu_{22} \\
&= P(n_{11}, n_{12}, n_{21}, n_{22})n_{12}\mu_{12} \qquad (1.2)
\end{aligned}$$

$$\begin{aligned}
&P(n_{11}+1, n_{12}, n_{21}-1, n_{22})(n_{11}+1)\mu_{11}p_{1,1,2,1} \\
&= P(n_{11}, n_{12}, n_{21}, n_{22})(n_{21}/(n_{21}+n_{22}))\mu_{21} \qquad (2.1)
\end{aligned}$$

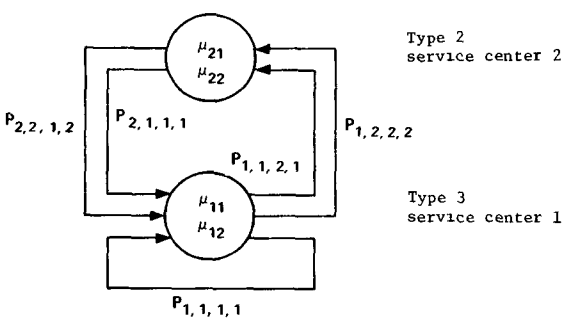

FIG. 2 Example network model

$$P(n_{11}, n_{12} + 1, n_{21}, n_{22} - 1)(n_{12} + 1)\mu_{12} = P(n_{11}, n_{12}, n_{21}, n_{22})(n_{22}/(n_{21} + n_{22}))\mu_{22}. \quad (2.2)$$

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. Independent 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 independent balance equations. Thus a solution for the independent balance equations is automatically a solution to the global balance equations. In many cases the independent balance equations are inconsistent and therefore have no solution. For example, if there is FCFS scheduling at a service center and different classes of customers have different service time distributions, the independent balance equations are inconsistent.

The value of the independent balance technique 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 many interesting cases do not have known solutions).

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

For each ergodic subchain E_k we define the following set of equations:

$$\sum_{(i,r)\in E_k} e_{ir}p_{i,r,j,s} + q_{js} = e_{js}, \quad (j, s) \in E_k.$$

The value of q_{js} is determined by the rate of exogenous arrivals of class s customers to service center j. If $q_{js} = 0$ $\forall(j, s) \in E_k$, then the network is closed with respect to E_k. In this case the e_{ir} are determined to within a multiplicative constant. e_{ir} can be interpreted as the relative arrival rate of class r customers to service center i. If not all of the $q_{js} = 0$ for $(j, s) \in E_k$, then we assume a unique solution for the e_{ir}. In this case e_{ir} is the absolute arrival rate of class r customers to service center i.

Note that a system may be "open" with 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 ith service center the rth class of customers has a service time distribution that is represented as a network of stages, then this is represented as shown in Figure 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 $A_{irl} = \prod_{j=1}^{l} a_{irj}$.

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

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

where C is a normalizing constant chosen to make the equilibrium state probabilities sum to

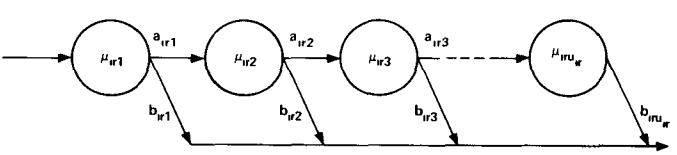

Fig. 3. Representation of the service time distribution of a class r customer at service center i

1, $d(S)$ *is a function of the number of customers in the 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 $f_i(x_i) = (1/\mu_i)^{n_i} \prod_{j=1}^{n_i} [e_{ix_{ij}}]$.

If service center i is of type 2, then $f_i(x_i) = n_i! \prod_{r=1}^{R} \prod_{l=1}^{u_{ir}} \{[e_{ir}A_{irl}/\mu_{irl}]^{m_{irl}}(1/m_{irl}!)\}$.

If service center i is of type 3, then $f_i(x_i) = \prod_{r=1}^{R} \prod_{l=1}^{u_{ir}} \{[e_{ir}A_{irl}/\mu_{irl}]^{m_{irl}}(1/m_{irl}!)\}$.

If service center i is of type 4, then $f_i(x_i) = \prod_{j=1}^{n_i} [e_{ir_j}A_{ir_jm_j}(1/\mu_{ir_jm_j})]$.

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

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

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

4. *Marginal Distributions*

The solutions presented in Section 3 for equilibrium state probabilities are in terms of states which contain more information than is usually required. For example, the ordering of customers in type 1 and type 4 service centers is part of the specification of a state. The more detailed states are necessary to derive the equilibrium state probabilities. In this section we exhibit some marginal distributions obtained by aggregating states. These marginal distributions are of interest because they lead to computationally more efficient means of calculating the normalization constant for closed networks and because of their implications.

4.1. Marginal Distributions and Their Implications. We define an aggregate system state as the number of customers of each class in each service center. More formally, an aggregate state S of the system is given by $(y_1, y_2, \cdots, y_N)$, where $y_i = (n_{i1}, n_{i2}, \cdots, n_{iR})$ and n_{ir} 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 $1/\mu_{ir}$ be the mean service time of a class r customer at service center i. Then the equilibrium state probabilities are given by

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

where

if service center i is of type 1, then $g_i(y_i) = n_i!\{\prod_{r=1}^{R} (1/n_{ir}!)[e_{ir}]^{n_{ir}}\}(1/\mu_i)^{n_i}$;

if service center i is of type 2 or 4, then $g_i(y_i) = n_i! \prod_{r=1}^{R} (1/n_{ir}!)[e_{ir}/\mu_{ir}]^{n_{ir}}$;

if service center i is of type 3, then $g_i(y_i) = \prod_{r=1}^{R} (1/n_{ir}!)[e_{ir}/\mu_{ir}]^{n_{ir}}$.

In each case the expression for $g_i(y_i)$ is derived by summing $f_i(x_i)$ over all x_i with $n_{i1}, n_{i2}, \cdots, n_{ik}$ 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 service rate at center i is the same for each class of jobs but depends on the number of customers at the center, then the factor $\prod_{r=1}^{R} (1/\mu_{ir})^{n_{ir}}$ is replaced by $\prod_{j=1}^{n_i}(i/\mu_i(j))$, where $\mu_i(j)$ is the service rate at service center i when there are j customers at this service center. Modifications to the solutions required by service rates that depend on the number of customers at a center are discussed in Section 5.

The implications of this result are clear. Although we began with almost general service time distributions for type 2, 3, and 4 service centers, only the mean service times

appear in $P(S = (y_1, y_2, \cdots, y_N))$. Thus for the aggregate states and within the bounds of the assumptions of the model, any service time distributions for the different classes of customers yield the same results as exponential service time distributions. It is important to note that while only the means of the service time distribution appear in the results, the effects of the different classes of customers is still present, i.e. the means $\{1/\mu_{ir}\}$ appear in the solution.

We note also that the normalization constant C can be more efficiently calculated from the aggregate states since there are fewer of the aggregate states.

4.2. Marginal Distributions for Open Systems. 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 paragraphs develop this simplification.

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

If service center i is of type 1, then $h_i(n_i) = (\sum_{r\in R_i} e_{ir})^{n_i}(1/\mu_i)^{n_i}$.

If service center i is of type 2 or 4, then $h_i(n_i) = (\sum_{r\in R_i} (e_{ir}/\mu_{ir}))^{n_i}$.

If service center i is of type 3, then $h_i(n_i) = (1/n_i!)(\sum_{r\in R_i} (e_{ir}/\mu_{ir}))^{n_i}$.

The evaluation of the normalizing constant requires summing the given expression for the equilibrium state probabilities over all feasible states. The simple recursive technique used by Buzen [4] extends to general networks with one class of customers. We now show a closed form solution for C for an open network.

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 $\lambda(M(S)) = \lambda = \text{constant}$. Since the system is open, any number of customers is feasible at a service center. Therefore

$$C^{-1} = \sum_{n_1=0}^{\infty}\sum_{n_2=0}^{\infty}\cdots\sum_{n_N=0}^{\infty}\left(\prod_{i=1}^{N}\lambda^{n_i}h_i(n_i)\right) \quad \text{or} \quad C^{-1} = \left(\sum_{n_1=0}^{\infty}\lambda^{n_1}h_1(n_1)\right)\left(\sum_{n_2=0}^{\infty}\lambda^{n_2}h_2(n_2)\right)\cdots\left(\sum_{n_N=0}^{\infty}\lambda^{n_N}h_N(n_N)\right).$$

Also,

$$\sum_{n_i=0}^{\infty} h_i(n_i) = \begin{cases}(1 - \sum_{r\in R_i} \lambda(e_{ir}/\mu_i))^{-1} & \text{if service center } i \text{ is type 1;}\\ (1 - \sum_{r\in R_i} \lambda(e_{ir}/\mu_{ir}))^{-1} & \text{if service center } i \text{ is type 2 or 4;}\\ \exp\left[\sum_{r\in R_i} \lambda(e_{ir}/\mu_{ir})\right] & \text{if service center } i \text{ is type 3.}\end{cases}$$

Note that the normalization constant factors into terms where each term involves only the parameters for a single service center. It follows that the equilibrium state probabilities factor into terms where each term involves only the parameters for a single service center. From this it is easily seen that the number of customers in each service center are independent random variables.

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

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

$\rho_i = \sum_{r\in R_i} \lambda(e_{ir}/\mu_{ir})$ 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;

$P_i(n_i) = e^{-\rho_i}(\rho_i^{n_i}/n_i!)$ 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 centers 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 $M/G/\infty$ system with $\rho_i = \lambda/\mu$. 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.

5. *State Dependent Service Rates*

Various forms of state dependent service rates can easily be incorporated into the network models The most straightforward case is when the service rate at a service center depends on the total number of customers at that service center.

Let $x_i(n_i)$ be an arbitary but positive function of the number of customers n_i at the ith service center. $x_i(n_i)$ is the rate of service at the ith service center when there are n_i customers at that service center *relative* to the service rate when $n_i = 1$. (Thus $x_i(n_1) = 1$.) With this type of state dependent service rate at service center i, $f_i(x_i)$ becomes $f_i(x_i)(1/\prod_{a=1}^{n_i} x_i(a))$. This form of state dependent service rate is useful, for example, when the ith service center contains multiple servers. If there are k_i servers then we might let

$$x_i(n_i) = \begin{cases} n_i, & 1 \leq n_i \leq k, \\ k, & n_i > k. \end{cases}$$

A case of multiple servers where the $x_i(n_i)$ function might be chosen differently occurs when the servers are central processors. To approximate the effect of memory interference, $x_i(n_i)$ would be less than n_i even when $n_i \leq k$.

Another form of state dependent service rates occurs when the service rate of a class r customer at service center i depends on the number n_{ir} of class r customers at service center i. This form of state dependent service rate cannot be modeled for type 1 service centers! Let $y_{ir}(n_{ir})$ be an arbitrary positive function of n_{ir} which is the service rate of

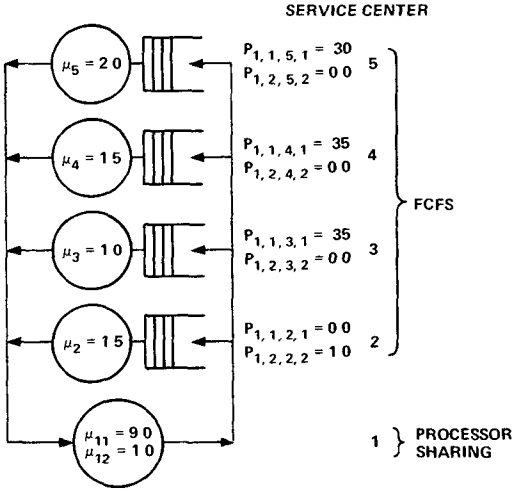

FIG 4. Example network model

class r customers at service center i relative to the service rate when there is one class r customer at service center i. In this case $f_i(x_i)$ is replaced by $f_i(x_i)\prod_{r=1}^{R}\prod_{a=1}^{n_{i,r}}(1/y_{i,r}(a))$.

A third form of state dependent service rates involves the number of customers in several service centers. Let $I = \{i_1, i_2, \cdots, i_m\}$ be a subset of the service centers. Let $n_I = \sum_{i\in I} n_i$ and let $Z_I(n_I)$ be an arbitrary positive function which is the relative rate of service to customers in the subset I of service centers relative to the service rates when n_I is one. In this case $\prod_{i\in I} f_i(x_i)$ becomes $\prod_{i\in I} f_i(x_i)\prod_{a=1}^{n_I}(1/Z_I(a))$.

Finally, we note that these various forms of state dependent service rates can be combined. For example, consider a subset I of service centers where the service rate at each service center $i \in I$ is a function of the number of customers n_I and n_i. In this case $\prod_{i\in I} f_i(x_i)$ becomes

$$\left\{\prod_{i\in I}\left((f_i(x_i)\prod_{a=1}^{n_i}(1/x_i(a))\right)\right\}\prod_{b=1}^{n_I}(1/Z_I(b)).$$

6. *Examples*

In this section we give simple examples that illustrate some of the results of the paper.

Example 1. Consider the system shown in Figure 4. 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 a type 2 center. 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, the rate at which class 1 customers leave service center 1, and r_2, the rate at which class 2 customers leave service center 1.

Now the equivalent customers have parameters given by

$$1/\mu_1 = r_1/(r_1 + r_2)1/\mu_{11} + (r_2/r_1 + r_2))(1/\mu_{12});$$

$$p_{1,i} = (r_1/(r_1 + r_2))p_{1,1,i,1} + (r_2/(r_1 + r_2))p_{1,2,i,2},\ i = 2, 3, 4, 5.$$

	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)

	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)

FIG. 5. (a) Two classes of customers; n_2 = number of class 2 customers = 1. (b) Same system with one class of equivalent customers, number of customers = $n_1 + n_2$

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 Figure 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 carries a prime and is 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 1 (the CPU) goes down initially as the number of equivalent customers increases from 1 to 2, and the difference for this server is substantial (between 4.5 and 9%). 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 when the system contains two or more customers.

Example 2. The customer class change concept can also be used to capture some complex sequencing properties of the system being modeled. For example, one of Moore's [13] models of a time-sharing system included a swapping drum. A simplified model is shown in Figure 7.

If a customer at the swapping drum has just come from the terminals, then it is being swapped into main memory and should next move to the CPU. If the customer has just come from the CPU, then it is being swapped out and should next move to the terminals.

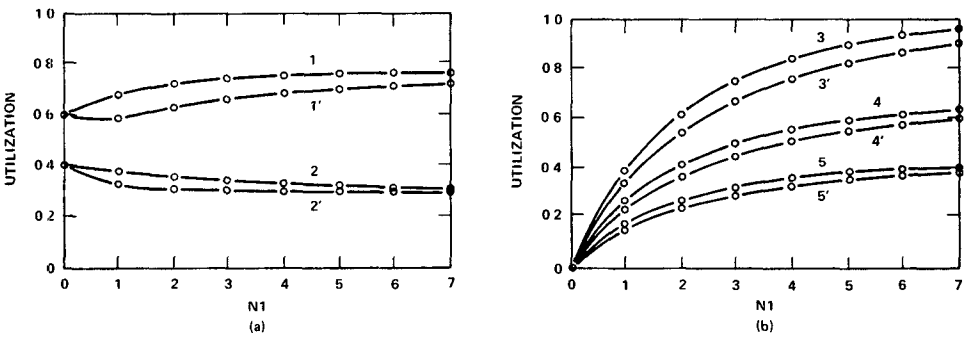

FIG. 6. Utilization of service centers versus number of customers for (a) different customers and (b) equivalent customers

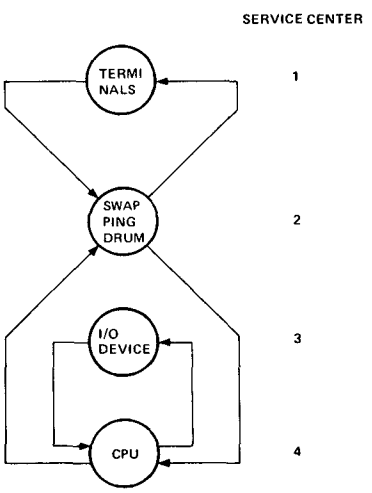

FIG. 7. A time-sharing system model with a swapping drum

However, without the concept of class changes, we can only define transition probabilities from the swapping drum to the CPU and from the swapping drum to the terminals. The natural selection would be to assign the value $\frac{1}{2}$ to each of these transition probabilities. It is easy to see that this is not an accurate model of the sequencing pattern. Moore did not have solutions available which allowed class changes and therefore used an approximation. In this approximation a customer made only one visit to the swapping drum between visits to the terminals, and the swapping drum service time was doubled. Using class changes we can model the actual sequencing. Customers can be in class 1 or class 2. Customers at the terminals are in class 1 and remain in class 1 when they move to the swapping drum. Class 1 customers move to the CPU from the swapping drum with probability 1 and remain in class 1. When leaving the CPU there is a probability of going to the I/O device or to the swapping drum. The transition from the CPU to the swapping drum is defined to be a change from class 1 to class 2 also. Class 2 customers leaving the drum have probability 1 of visiting the terminals next. The transition probabilities are more formally described as

$$p_{1,1,2,1} = 1, \quad p_{2,1,4,1} = 1, \quad p_{4,1,3,1} + p_{4,1,2,2} = 1, \quad p_{3,1,4,1} = 1, \quad p_{2,2,1,1} = 1.$$

The class of models in [12] also allows representation of complex sequencing properties.

7. *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 with state independent arrivals, 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 1 service centers (FCFS scheduling) are appropriate models of secondary storage I/O devices because preemptive scheduling is usually not possible or efficient for such devices. Type 2 and type 4 service centers (processor-sharing scheduling and LCFS) are appropriate models for CPUs since LCFS is an efficient preemptive scheduling method and round robin scheduling approaches processor sharing; both have been found to improve the performance of CPUs. Type 3 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 frequent objections to queueing models as models of computer systems. The examples given indicate how significant different classes of customers can be in the utilization levels predicted by model analysis and in the systems captured by models.

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.

REFERENCES

(Note References [8, 14] are not cited in the text.)

1. Baskett, F The dependence of computer system queues upon processing time distribution and central processor scheduling. Proc ACM-SIGOPS Third Symposium on Operating System Principles, Stanford U., Stanford, Calif., Oct. 1971, pp. 109–113.

2. BASKETT, F., AND PALACIOS, F G. Processor sharing in a central server queueing model of multiprogramming with applications. Proc. Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton U , Princeton, N J. March 1972, pp. 598–603.
3. BUZEN, J. Queueing network models of multiprogramming. Ph.D. Th., Div. of Eng. and Appl. Sci., Harvard. U., Cambridge, Mass , 1971.
4. BUZEN, J. Analysis of system bottlenecks using a queueing network model. Proc. ACM-SIGOPS Workshop on System Performance Evaluation, April 1971, pp. 82–103.
5 CHANDY, K. M The analysis and solutions for general queueing networks Proc. Sixth Annual Princeton Conference on Information Sciences and Systems, Princeton U., Princeton, N. J., March 1972, pp. 224–228.
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. Ninth Annual Design Automation Workshop, Dallas, Tex., June 1972, pp 151–153.
7. COX, D. R. A use of complex probabilities in the theory of stochastic processes. *Proc. Cambridge Phil. Soc. 51* (1955), 313–319
8. FELLER, W. *An Introduction to Probability Theory and Its Applications, Vol. 1*, 3rd ed Wiley, New York, 1968.
9. FERDINAND, A. E. An analysis of the machine interference model *IBM Syst. J. 10*, 2 (1971), 129–142.
10. GORDON, W J , AND NEWELL, G F. Closed queueing systems with exponential servers *Oper Res 15* (1967), 254–265
11. JACKSON, J R Jobshop-like queueing systems *Manage. Sci 10*, 1 (Oct 1963), 131–142
12. LAVENBERG, S. S Queueing analysis of a multiprogrammed computer system having a multilevel storage hierarchy. *SIAM J Comput 2*, 4 (1973), 232–252.
13. MOORE, C G , III Network models for large-scale time-sharing systems Tech. Rep. No. 71-1, Dep. of Ind Eng , U of Michigan, Ann Arbor, Mich., April 1971
14. PALACIOS, F G An analytic model of a multiprogramming system including a job mix Rep TR-4, Dep. of Computer Sciences, U. of Texas at Austin, Austin, Tex., June 1972.
15 POSNER, M., AND BERNHOLTZ, B Closed finite queueing networks with time lags and with several classes of units *Oper Res. 16* (1968), 977–985.
16. SAKATA, M., NOGUCHI, S., AND OIZUMI, J Analysis of a processor-shared queueing model for time-sharing systems Proc Second Hawaii International Conference on System Sciences, U. of Hawaii, Honolulu, Hawaii, Jan 1969, pp 625–628.
17 WHITTLE, P Nonlinear migration processes. *Internat. Statist Inst Bull 42*, Bk. 1 (1969), 642–647.
18 WHITTLE, P. Equilibrium distributions for an open migration process. *J Appl. Probabil. 5* (1968), 567–571

RECEIVED AUGUST 1972; REVISED AUGUST 1974