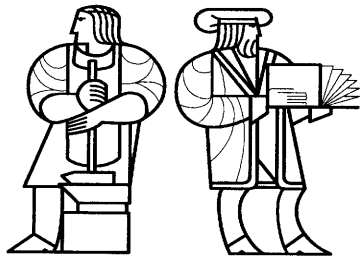
OPERATIONS RESEARCH CENTER

Working Paper

An N Server Cutoff Multi-Priority Queue

by
C. Schaack and
R. C. Larson

OR 135-85 February 1985



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

An N Server Cutoff Multi-Priority Queue

Christian Schaack
Richard C. Larson

OR 135-85 February 1985

The research for this report was funded in part by the National Institute of Justice on grants no. 83-IJ-CX-0065 and 84-IJ-CX-0063 and by the National Science Foundation on grant no. 8411871-SES.

ABSTRACT

Consider a multi-priority, nonpreemptive, N-server Poisson arrival queueing system. Service times are negative exponential. In order to save available servers for higher priority customers, arriving customers of each lower priority are deliberately queued whenever the number of servers busy equals or exceeds a given priority-dependent cutoff number. A queued priority i customer enters service the instant there are fewer than the respective cutoff number of servers busy and all higher priority queues are empty. The principal result is the priority i waiting time mean, second moment, and distribution (in transforms). The analysis is extended to systems in which any subset of priority levels may overflow to some other system, rather than join infinite capacity queues. The paper concludes with illustrative computational results.

1 INTRODUCTION

We consider a multi-server queueing system with different priority classes of customers. Arrivals in each priority stream occur in a random (i.e., Poisson) manner, and each customer's service requirements are governed by a negative exponential distribution whose mean is independent of priority class. Associated with each priority class i there is a "server cutoff number" C_i, where $i=1,2,...,$ $T \equiv$ number of different priority classes. If a priority i customer finds upon arrival fewer than C_i servers busy, then she enters service immediately. Otherwise (i.e., if C_i or more servers are busy upon arrival), the customer joins an infinite capacity first-in-first-out (FIFO) queue of other priority i customers. In order to depict the relative urgencies of the respective priority levels, we assume the C_i's are ordered so that $C_T \leq \cdots \leq C_2 \leq C_1$. If $C_{i+1} < C_i$, the priority i queue, if nonempty, is reduced by one upon the instant that fewer than C_i servers are busy. If two or more priority streams have the same server cutoff number (i.e., if $C_{i+1} = C_i$), then all the higher priority customers are served in a head-of-the-line manner before any of the lower priority customers. (By convention, $i=1$ designates "highest priority.")

The model is motivated by applications in police and ambulance dispatching, hospital bed management, communications channel allocation, and any other priority queueing system in which it is desirable to retain a "strategic reserve" of servers for higher priority customers. The current practice of most police departments, for instance, is to deplete the pool of available servers (i.e., police patrol cars) to zero before delaying any customers ("police calls for service") in queue. Yet it is well known (Larson [1972], Tien [1976]) that in most police departments there are from three to seven different priority classes of police calls for service and only the most urgent (e.g., felony in progress, officer in trouble) require immediate service. Our model, when exercised with typical police operational data and

reasonable delay cost structures, reveals the desirability of deliberately delaying lower priority customers in queue, even while available servers are present, as an "insurance" against near-term more urgent customers who may arrive.

We derive the following performance measures for the system, assumed to be operating in steady state: mean and higher moments of the delay in queue experienced by priority i customers; probability that a priority i customer experiences no queueing delay; server utilization factor. We also generalize our results to allow any particular priority classes to be "lost to the system" when arrivals find too many servers busy, rather than to enter a queue. In applying the model, we determine optimal server cutoffs (C_i) under alternative cost structures.

Our analysis of the system hinges on developing an iterative scheme for determining the probability distribution for the number of busy servers. By starting at the highest priority level ($i=1$), and continuing through each successively lower priority level, we replace at each priority level the actual multi-server queue with a fictitious but equivalent single server queue with general service time. This replacement is motivated by the fact that each queue (priority $i=1,2,...$) acts when nonempty as if it were a Poisson arrival, single server, general service time queue. Given this observation, the analysis utilizes standard busy period and semi-Markov methods.

In Sections 2 and 3, respectively, we define the model and review the related literature. The detailed analysis is developed in Section 4, with several derivations relegated to an appendix. Extensions to loss type systems are developed in Section 5, and illustrative computational results are given in Section 6.

2 MODEL DESCRIPTION

In this section we provide details of the basic model, which assumes that arriving customers either enter service immediately or join a priority-specific infinite capacity FIFO queue. Section 5 extends this model to loss systems.

Customers are assumed to arrive in a homogeneous Poisson manner to an N server queueing system, with arrival rate λ_i (customers/unit time) for priority i customers ($i = 1,2,...,T$). All Poisson streams operate independently. By convention, type i customers have higher priority than type j customers if $i<j$. Service time is assumed to be negative exponential with mean $1/\mu$, independent of the priority of the customer or the identity of the server.

The service discipline is assumed to be non-preemptive. Priority i customers enter service immediately upon arrival only if there are less than C_i servers busy. Otherwise they are backlogged in a queue of other priority i customers; this queue is depleted in a FIFO manner, with each depletion instant corresponding to a moment of service completion arising when precisely C_i servers are busy. If $C_{i-1} = C_i$, then the priority i-1 queue must be empty before priority i customers are serviced (HOL). By convention, the server cutoff number for the highest priority customers is $C_1 = N$, the number of servers.

A proposed shorthand notation for our model is $M/M/\{C_i\}$, designating Markovian (Poisson) input, Markovian (negative exponential) service times, and a set of server cutoffs $\{C_i\}$. The model is summarized as follows:

- N identical servers
- T priority levels of customers
  λ_i = Poisson arrival rate of type i customers, $i = 1,...,T$
  μ = exponential service rate (identical for all priority levels)
- Type i customers enter service immediately upon arrival only if less than C_i servers are busy, where $0 < C_T \leq C_{T-1} \leq ... \leq C_2 \leq C_1 = N$; otherwise they join a

FIFO queue of other priority i customers, the queue being depleted at instants of service completion arising when precisely C_i servers are busy. If two or more adjacent priority levels have the same server cutoff number, then higher priority customers are always served before any queued lower priority customers (i.e., the discipline is HOL by priority.)

3 LITERATURE REVIEW

The two priority case, namely M/M/{N,C_2}, has been studied by various researchers since 1966, when it was first formulated by Benn [1966] who applied it to a railroad transportation problem. A detailed solution of this two-priority problem was published by Jaiswal [1968, p. 204ff.], who analyzes the following variants: high and low priority calls queued; high priority calls queued, low priority calls lost; and, high priority calls lost, low priority calls queued. The analysis proceeds from the steady state balance equations. Using a fairly involved discrete transform technique, Jaiswal obtains the steady state probabilities for the number of busy servers; unfortunately, these results are in a very inconvenient form. Other researchers have used the same model in various applications, mostly in health care planning: Shonick and Jackson [1973], Abol'nikov and Yasnogorodskiy [1974], McClain [1976], Esogbue and Singh [1976].

Our own work was largely motivated by a recent paper by Taylor and Templeton [1980] who improved upon the solution for the two-priority problem published in Jaiswal. Paralleling the derivation in Jaiswal [1968], Taylor and Templeton set up the steady state global balance equations for states defined by (n, q_1, q_2), where n is the number of busy servers; q_1, the number of high priority customers queued; and q_2, the number of low priority customers queued. In order to obtain the steady state probabilities, they take z-transforms of the balance equations with respect to the variable q_2. Then, departing from Jaiswal, they proceed to solve for
$P_n \equiv$ Probability that precisely n servers are busy (n=0,1,...,N), by inversion of a band matrix. The procedure is recursive and fairly involved, but the results are interesting and remarkably simple.

The steady state probabilities for the number of busy servers are given by

$$P_n = P_o \frac{\rho^n}{n!} \qquad \text{for } 0 \leqq n < C_2$$

$$P_n = P_o \frac{\rho^{C_2} \rho_1^{n-C_2}}{n!} \frac{C_2}{C_2 - \rho_2 S(C_2)} \qquad \text{for } C_2 \leqq n < C_1 = N$$

$$P_n = P_o \frac{\rho^{C_2} \rho_1^{N-C_2}}{N!} \frac{C_2}{C_2 - \rho_2 S(C_2)} \frac{N}{N-\rho_1} \qquad \text{for } n = C_1 = N$$

where

$$\rho_1 = \frac{\lambda_1}{\mu}, \quad \rho_2 = \frac{\lambda_2}{\mu}, \quad \rho = \rho_1 + \rho_2 ,$$

$$S(j) = \rho_1^j j! \left[\sum_{i=j}^{N-1} \frac{\rho_1^i}{i!} + \frac{\rho_1^N}{N!} \frac{N}{N-\rho_1} \right] \qquad \text{for } C_2 \leqq j \leqq N-1,$$

and P_0 is obtained by normalization.

No interpretation of S(j) is given by Taylor and Templeton [1980] and there is no further investigation into the apparent structure of the results. Although the final result appears to be a closed form expression, it really is not: $S(C_2)$ has to be computed recursively. Given the complexity of the derivation, there is little hope of extending the results to more than two priorities using the Taylor and Templeton method. Taylor and Templeton [1980] also derive waiting time distributions and discuss some variants of the two-priority problem. While these results are most interesting indeed, we should stress the lack of physical intuition emerging from them. The "closed form" of the low priority waiting time transform, for example, is extremely complicated. We believe it would be better to leave the results under a recursive form in terms of quantities that have a physical meaning. This we shall endeavor to do in Section 4 when we analyze the T-priority cutoff model. The insight

we shall gain into the structure of the problem will help us solve it by inspection for any number of priority classes.

Finally, to conclude the literature review, there has been little work done on the T-priority problem ($T \geq 3$) to the authors' knowledge. Cooper [1972] mentions an unpublished paper by Descloux that seems to have taken a closer look at some aspects of the problem. The authors have been unable to obtain a copy of this paper. Cobham's [1954] well-known head-of-the-line T-priority problem has no server cutoffs; hence, in our notation, Cobham's model is a queue of type $M/M/\{C_i = N : i = 1,2,...,T\}$. We shall derive Cobham's results as a special case.

4 ANALYSIS

4.1 The M/G/1 Approach

This paper extends the results of the model described above to more than two priorities by following a probabilistic approach based on M/G/1 queueing theory. Contrary to Jaiswal [1968] and Taylor and Templeton [1980], we do not work with the global balance equations directly. Our method is based on continuous as opposed to discrete transforms. This approach yields a physical understanding of the problem.

For the analytical developments of the following sections, it is helpful to think of the queueing system in the following way: Suppose customers of priority i are waiting in queue i and have no information about the queues of other priorities, which form in other waiting rooms of the service facility (Figure 1). For all they know, they may be in the only queue in the system. Assume that the customers in queue i can only observe how their own queue behaves, i.e., when the next customer in their queue begins service. Then, the following is true: Either a new priority i arrival finds no queue and a free server on arrival (i.e., fewer than C_i servers are busy) and enters service immediately; or the arrival finds the system busy for her purpose, (i.e., at least C_i servers are busy), in which case she joins the end of a queue of other priority i customers; she observes that the times between successive "move ups" in queue position (say from position k to k-1, $k \geq 1$) are independent, identically distributed (i.i.d.) with a general distribution for the time between move ups. This queueing behavior is identical to that of an M/G/1 system, where G depicts a general service time distribution represented here by the time between successive move ups. G is not, however, the distribution of time actually spent in service by a type i customer; that distribution remains M (negative exponential) with mean $1/\mu$. The

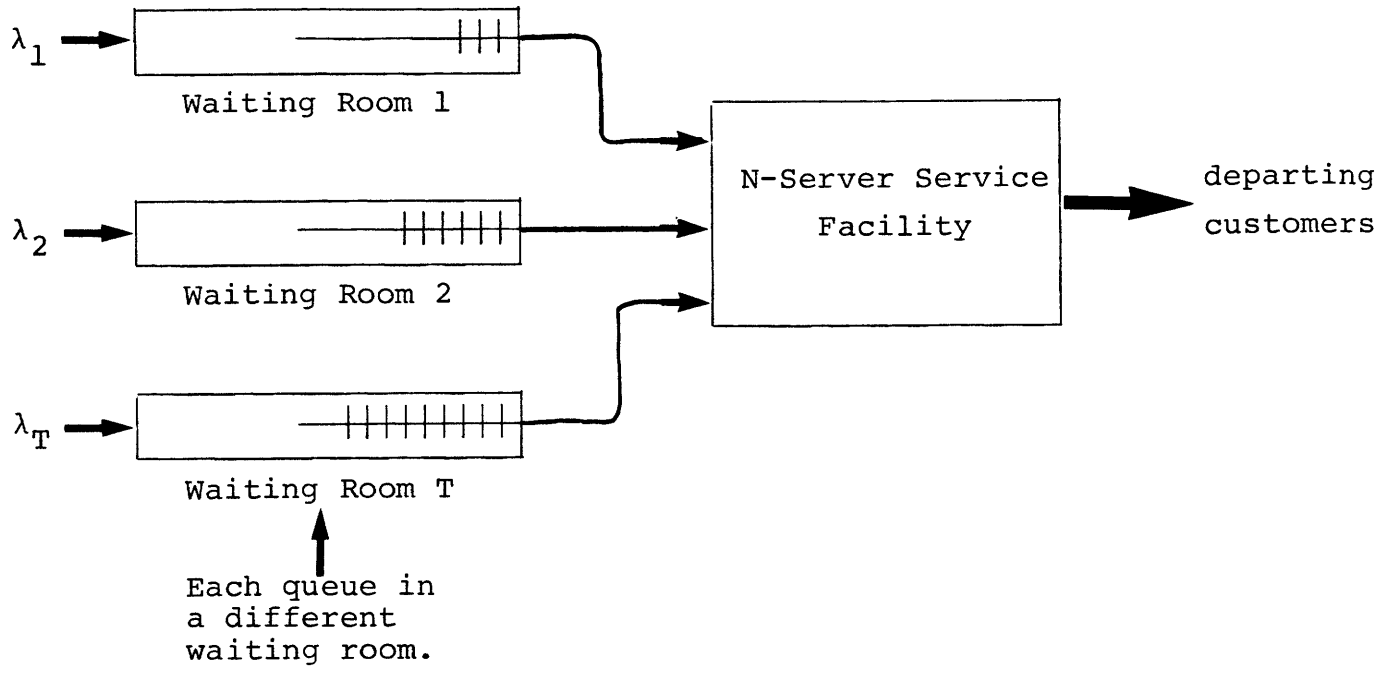


Figure 1

Arrival Streams of Different Priority

Enter Separate Queues (waiting rooms)

observed G for times between move ups is in fact the probability distribution of a busy period ("busy" for class i) induced by higher priority arrivals (i.e., of priorities 1,2,...,i-1), whose existence she is unaware of.

This, then, is how we shall proceed in Section 4: Determine the "general" service time distribution seen by a queued priority i customer. As we shall see, it is then easy to derive the probabilities P_n that n servers are busy (for $n=0, ...,N$), as well as the waiting time distributions (in transform domain) for the various priorities.

4.2 The Distribution of the Number of Busy Servers

To get a better understanding of the structure of the problem, it is helpful to look at the three-dimensional state transition diagram of the (two-priority) cutoff problem (Figure 2). (The two-dimensional view of the state space (of, e.g., Jaiswal [1968 p. 209]) hides some of the geometric structure.) Let $\Pi(n, q_1, q_2)$ be the steady state probability that the system is in state (n, q_1, q_2); then

$$P_n = \sum_{q_1,q_2} \Pi(n,q_1,q_2)$$

Figure 2 shows that the P_n's, the probabilities that there are n servers busy, represent the total probability of being in a (hyper-)plane orthogonal to the n-axis and passing through (n,0,0). Note that this result holds for T priorities, ($T \geq 2$). Indeed then, the state space representation analogous to Figure 2 is $(T+1)$-dimensional and P_n is the probability of being in a T-dimensional hyperplane orthogonal to the n-axis and passing through (n,0,0,...,0).

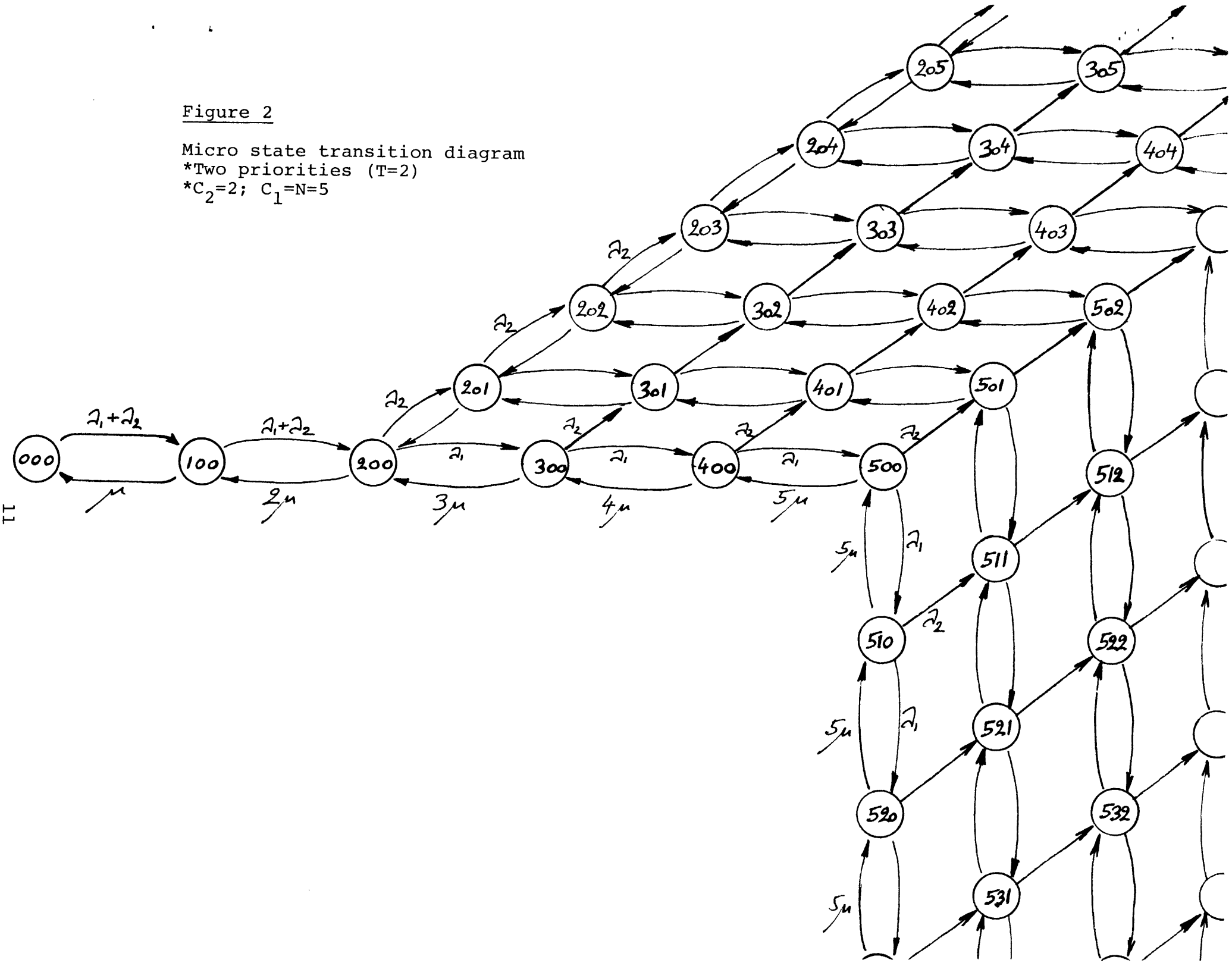
From here on, we shall concern ourselves with the general T-priority problem. We shall no longer be interested in the micro-states $(n, q_1, q_2,...,q_N)$, but rather we shall work with a "macro-state space" where the (macro-)states are $\{S_n\}$, the number of busy servers, and P_n, the corresponding steady state probabilities.

Using "balance-of-flow" results from M/M/N queues, it is easy to see that

$$P_n = P_{n-1}\frac{\rho}{n} \text{ for } 0<n<C_T, \text{ where } \rho \equiv \sum_{j=1}^{T}\frac{\lambda_j}{\mu} \quad . \tag{1}$$

Figure 2

Micro state transition diagram
*Two priorities (T=2)
*C_2=2; C_1=N=5



Similarly, it is easy to see by summing the global balance equations over q_i (see Figure 2) that

$$P_n = P_{n-1}\frac{\rho_i^c}{n} \text{ for } 1 \leq i < T \text{ and } C_{i+1} < n < C_i, \quad \text{where } \rho_i^c \equiv \sum_{j=1}^{i}\frac{\lambda_j}{\mu} \quad . \tag{2}$$

Equations (1) and (2) show that, to obtain the steady state probabilities P_n, all that remains to be determined are the relationships between P_{C_i-1} and P_{C_i}, for i=1,...,T. Normalizing the sum of the P_n's to 1 then yields P_0.

In order to derive the probability distribution for the number of busy servers, we need to introduce some additional concepts and notation. Let the r.v. R_n denote the first passage time from S_n to S_{n-1} (n=1,2,...,N). Let the r.v. R_n^i be the first pasage time from S_n to S_{n-1} (n=1,2...,N) for a system with arrival streams of priority 1 through i only. (Figure 3 illustrates the definitions of the random variables R_n^i.) Due to the preferential service discipline enjoyed by higher priority customers, it should be clear that $R_n^{i-1} = R_n$ for all $n > C_i$, since S_{c_i} is that state at which priority i customers may be seen as being queued. If $C_i < C_{i-1}$ any service completion occurring from (macro-) state S_{c_i} will immediately result in either: (1) a reduction of the priority j queue by one, where j is the highest priority nonempty queue for which $C_j = C_i$, or: (2) cause a downward (macro-) state transition to state $S_{(c_{i-1})}$ (if all queues having cutoffs equal to C_i are empty). If $C_i = C_{i-1}$, then the priority i queue can be depleted only when the priority (i-1) queue is empty (i.e., the priority (i-1) queue empties before the priority i queue).

Focusing on the priority i queue, suppose the $(\ell+1)^{st}$ priority i customer waiting in the priority i queue becomes the ℓ^{th} customer in line at time τ_i and the $(\ell\text{-}1)^{st}$ customer in line at time τ_i' ($\ell \geq 1$); then $\tau_i' - \tau_i$ is defined to be the queue move up time B_i for priority i customers. Due to the Markovian nature of the entire system, it should be clear that the B_i's are i.i.d. (independent of ℓ, of course). Since the entry of a queued priority i customer into service leaves the system in (macro) state S_{c_i}, the

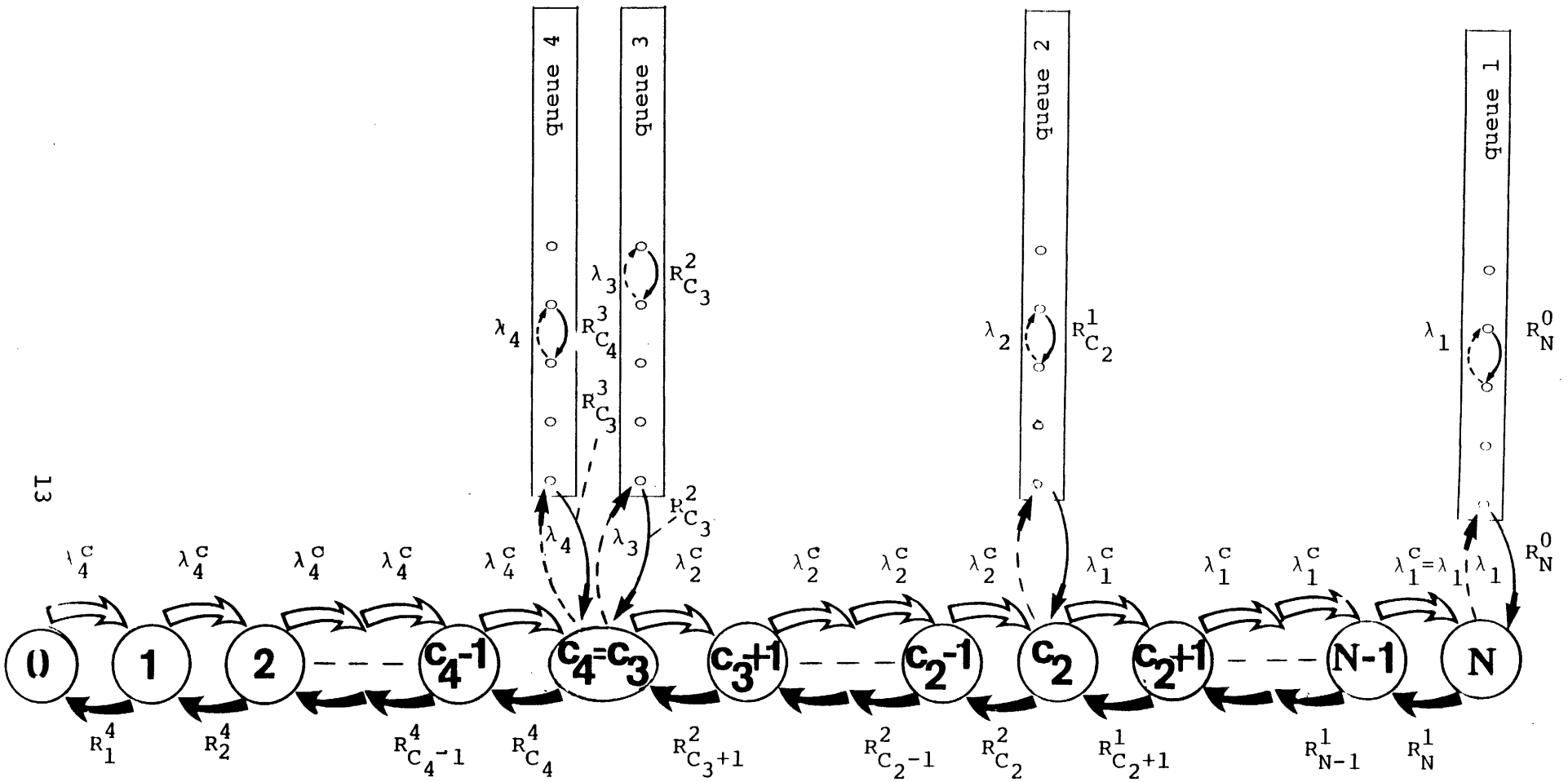


Figure 3

Illustration of the r.v.'s R^i_n, for $0 \le i \le T$ and $C_{i+1} \le n \le C_i$ for a four priority system with $0 < C_4 = C_3 < C_2 < C_1 = N$

*The black arcs show the first passage times R^i_n from one state to another.

The white arcs show the conditional upward transition rates for a given number of busy servers ($\lambda^c_i = \sum_{k=1}^{i} \lambda_k$)

next queued priority i customer (if any) will enter service (i.e., move up to position 0 in line) upon the next service completion arising from state S_{C_i}, assuming $C_i < C_{i-1}$. If $C_i = C_{i-1}$, then move ups in the priority i queue occur at moments of service completion arising from state S_{C_i} when queue i-1 is empty. The move up time is equivalent to a first passage time from S_{C_i} to $S_{C_{i-1}}$, assuming that no priority i (or lower priority) customers existed. Hence,

$$B_i = R_{C_i}^{i-1} \quad i=1,2,...,T, \tag{3}$$

defining $R_N^0 \equiv B_1$

We now have the ingredients for a recursive procedure to analyze the system. Derivation of the P_n's will require knowledge of the mean first passage times $E[R_n^i]$ for a certain combination of indices n and i; we will derive these quantities recursively starting with the highest priority customers.

Let the cumulative distribution function (cdf) of R_n^i be denoted by $R_n^i(y) \equiv \text{Prob}\{R_n^i \leq y\}$, with Laplace-Stieljtes transform

$$\tilde{R}_n^i(s) \equiv \int_0^\infty e^{-sy}\, dR_n^i(y) = E\left[e^{-sR_n^i}\right].$$

Starting with customers of priority class 1 (highest priority), it should be clear that any priority 1 customer waiting in queue experiences a queue move up time that is exponentially distributed with rate Nμ. Hence, recalling $B_1 \equiv R_N^0$,

$$\tilde{R}_N^o(s) = \frac{N\mu}{N\mu + s} \tag{4}$$

Note that the priority 1 queue moves (during a nonempty period) as if the N individual servers, each having service rate μ, were replaced by a single server having service rate Nμ.

Ignoring the existence of lower proirity customers, the first passage time R_N from S_N to S_{N-1} is a busy period for all N servers sustained by priority 1 customers. This busy period is equivalent to the busy period of an M/G/1 system having arrival rate λ_1 and negative exponential service time distribution with rate $N\mu$. Following the usual sub-busy period argument (see Kleinrock [1975]), we first condition on the duration of the first service, X_1, of the busy period in the equivalent M/G/1 system and on the number k of arrivals during X_1, then we can write

$$E\left[e^{-sR_N^1} \mid X_1 = y, K = k\right] = e^{-sy}\left[\tilde{R}_N^1(s)\right]^k$$

Deconditioning on k, we find

$$E\left[e^{-sR_N^1} \mid X_1 = y\right] = e^{-sy}\sum_{k=0}^{\infty}\frac{(\lambda_1 y)^k e^{-\lambda_1 y}}{k!}\left[\tilde{R}_N^1(s)\right]^k = e^{-y[s+\lambda_1-\lambda_1\tilde{R}_N^1(s)]}$$

Finally, removing the conditioning on X_1,

$$\tilde{R}_N^1(s) \equiv E\left[e^{-sR_N^1}\right] = \int_0^{\infty} E\left[e^{-sR_N^1} \mid X_1 = y\right] dX_1(y)$$

or,

$$\tilde{R}_N^1(s) = \tilde{R}_N^0\left(s+\lambda_1-\lambda_1\tilde{R}_N^1(s)\right)$$

If $C_2 < N$, then $R_N^1 = R_N$ and we must confront the problem of computing R_{N-1}. If $C_2 = N$, that is, if arriving priority 2 customers are also served immediately if at least 1 server is free, then by (3) $B_2 = R_N^1$ and a similar argument based on sub-busy periods applies. Generalizing, any priority i level queue can be seen to move (when nonempty) as an M/G/1 queue with arrival rate λ_i and service time (i.e., move up time) equal to the first passage time from (macro) state S_{c_i} to S_{c_i-1} for a system

having customers of priority 1 through i-1 only. Hence,

$$\widetilde{R}^{i}_{C_i}(s) = \widetilde{R}^{i-1}_{C_i}\left(s+\lambda_i-\lambda_i\widetilde{R}^{i}_{C_i}(s)\right) \qquad i=1,2,\ldots,T \tag{5}$$

A somewhat different (and easier) argument is required to obtain equivalent functional equations for the transforms of R^i_n for $C_{i+1} \leq n < C_i$. Given that there are precisely n servers busy ($C_{i+1} \leq n < C_i$), then the probability that the next transition in the number of servers is to $n+1$ is

$$r_n = \frac{\lambda^c_i}{\lambda^c_i + n\mu} \qquad where\, \lambda^c_i = \sum_{k=1}^{i} \lambda_k ,$$

and the probability that the next transition is to n-1 busy servers is $1-r_n$. Let the r.v. V_n represent the time until the next transition and let $\widetilde{V}_n(\,s\,)$ be the Laplace transform of its distribution; clearly V_n is exponentially distributed with rate $\lambda^c_i + n\mu$. Conditioning on the type of transition, we can write for R^i_n,

$$E\left[\, e^{-sR^i_n} \,\middle|\, upward\, transition \,\right] = \widetilde{V}_n(s)\widetilde{R}^{i}_{n+1}(s)\widetilde{R}^{i}_{n}(s)$$

and

$$E\left[\, e^{-sR^i_n} \,\middle|\, downward\, transition \,\right] = \widetilde{V}_n(s).$$

Removing the conditioning we find

$$\widetilde{R}^{i}_{n}(s) = r_n\widetilde{V}_n(s)\widetilde{R}^{i}_{n+1}(s)\widetilde{R}^{i}_{N}(s) + (1-r_n)\widetilde{V}_n(s),$$

or,

$$\widetilde{R}^{i}_{n}(s) = (1-r_n)\widetilde{V}_n(s)\left[\, 1-r_n\widetilde{V}_n(s)\widetilde{R}^{i}_{n+1}(s) \,\right]^{-1}$$

Substituting known values for r_n and $V_n(s)$, we find

$$\widetilde{R}^{i}_{n}(s) = n\mu\left[\, s+n\mu+\lambda^c_i-\lambda^c_i\widetilde{R}^{i}_{n+1}(s) \,\right]^{-1} \tag{6}$$

for $C_{i+1} \leq n < C_i$, $1 \leq i < T$.

Equations (4), (5), and (6) provide the results necessary for computing the P_n's. Straightforward differentiation of (4), (5), and (6) yields the following recursive

equations relating the first moments:

$$E\left[R_N^0 \right] = \frac{1}{N\mu} \tag{7}$$

$$E\left[R_{C_i}^i\right] = \frac{E\left[R_{C_i}^{i-1}\right]}{1-\lambda_i E\left[R_{C_i}^{i-1}\right]} \qquad i=1,2...T \tag{8}$$

$$E\left[R_n^i\right] = \frac{1}{n\mu}\left(1+\lambda_i^c E\left[R_{n+1}^i\right]\right) \qquad \begin{matrix} i=1,2,...,T-1 \\ C_{i+1}\le n<C_i \end{matrix} \tag{9}$$

Recall also that

$$E\left[B_i\right]=E\left[R_{C_i}^{i-1}\right] \qquad i=1,2...T$$

A flow chart illustrating the use of (7), (8) and (9) for computing all the relevant $E[R_n^i]$'s is given in Figure 4.

Armed with the above results, it is now easy to derive the relationships between the steady state probabilities P_{c_i-1} and P_{c_i}. For notational simplicity, define n^+ as the superstate, "At least n servers are busy." Let P_n denote the steady state probability corresponding to this superstate. Define f as the lowest priority class whose cutoff is C_i (i.e., $f=\max\{k:C_k=C_i\}$ and g as the highest priority whose cutoff is C_i (i.e., $g=\min\{k:C_k=C_i\}$). The system leaves state C_{i-1} for superstate C_i^+ with exponential rate λ_f^c. The holding time in superstate C_i^+ is given by the r.v. $R_{C_i}^f$. Now, applying Little's law locally to superstate C_i^+, we can write

$$P_{C_i}^+ = (\lambda_f^c P_{C_i-1}) E\left[R_{C_i}^f\right] \tag{10}$$

Of course, we can repeat this procedure with state C_i+1. Transitions to superstate $(C_i+1)^+$ (from state C_i) occur with conditional exponential rate λ_{g-1}^c and the holding time on $(C_i+1)^+$ is given by the r.v. $R_{C_i+1}^{g-1}$. Hence,

$$P_{C_i+1}^+ = \left(\lambda_{g-1}^c P_{C_i}\right) E\left[R_{C_i+1}^{g-1}\right] \tag{11}$$

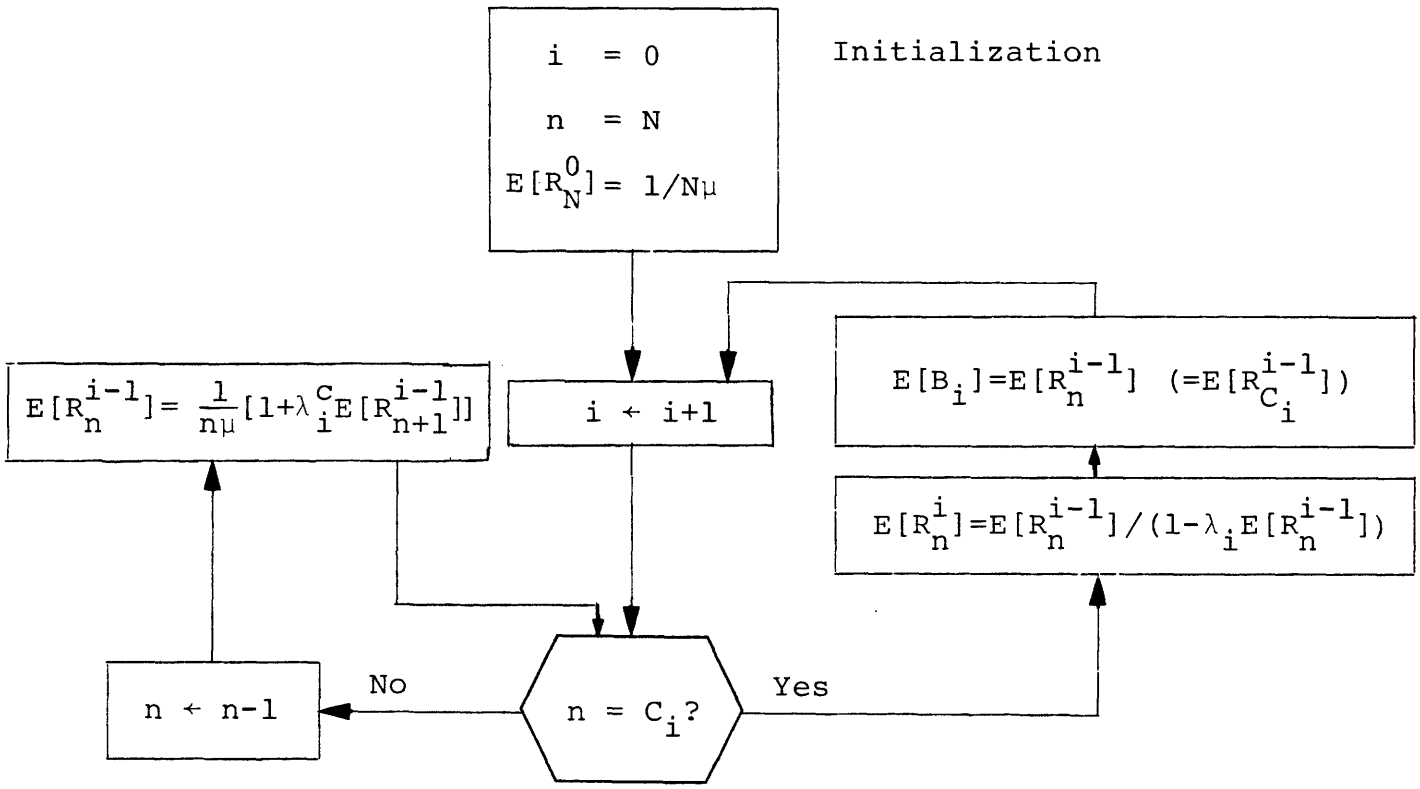


Figure 4

Flow Chart for Computing Mean First Passage Times and Mean Queue Moveup Times

Using the fact that $P^{+}_{c_i+1} = P^{+}_{c_i} - P_{c_i}$, we can write

$$\lambda^{c}_{g-1} P_{C_i} E\left[R^{g-1}_{C_i+1}\right] = \lambda^{c}_{f} P_{C_i-1} E\left[R^{f}_{C_i}\right] - P_i \quad ,$$

or,

$$P_{C_i} = P_{C_i-1} \frac{\lambda^{c}_{f} E\left[R^{f}_{C_i}\right]}{1+\lambda^{c}_{g-1} E\left[R^{g-1}_{C_i+1}\right]} \tag{12}$$

Using (8) and (9), (12) can be rewritten as

$$P_{C_i} = P_{C_i-1} \frac{E\left[R^{f-1}_{C_i}\right]}{E\left[R^{g-1}_{C_i+1}\right]} \frac{\rho^{c}_{f}}{C_i} \frac{1}{1-\lambda_f E\left[R^{f-1}_{C_i}\right]} \tag{13}$$

Applying (8) repeatedly, we can write

$$P_{C_i} = P_{C_i-1} \frac{\rho^{c}_{f}}{C_i} \prod_{k C_k = C_i} \left(\frac{1}{1-\lambda_k E\left[R^{k-1}_{C_i}\right]} \right) \tag{14}$$

This then concludes the derivation of the steady state probabilities, P_n, of the number of busy servers. Equations (2) and (14), repeated below, fully determine the P_n's:

$$P_n = P_{n-1} \frac{\rho^{c}_{i}}{n} \qquad \text{for } 1 \leq i < T, C_{i+1} \leq n < C_i \text{ and for } i = T, 0 \leq n < C_N \tag{2}$$

$$P_{C_i} = P_{C_i-1} \frac{\rho^{c}_{f}}{C_i} \prod_{k C_k = C_i} \left(\frac{1}{1-\lambda_k E\left[R^{k-1}_{C_i}\right]} \right) \qquad \text{for all } C_i\text{'s; where } f = \max\{k : C_k = C_i\} \tag{14}$$

We can now write down the P_n's by inspection, in terms of the expected service times seen by the various queues, where for the priority i queue the expected service time is $E[B_i] = E[R^{i-1}_{c_i}]$. For example, for a four-priority system with arrival rates λ_1, λ_2, λ_3, λ_4, individual server service rate μ, and cutoffs $0 < C_4 < C_3 = C_2 < C_1 = N$, the

steady state probabilities are as follows:

$$P_n = P_0 \frac{\left(\rho_4^c\right)^n}{n!} \qquad 0 \leqq n < C_4$$

$$P_n = P_0 \left(\frac{\rho_4^c}{\rho_3^c}\right)^{C_4} \frac{\left(\rho_3^c\right)^n}{n!} \frac{1}{1-\lambda_4 E[B_4]} \qquad C_4 \leqq n < C_3 = C_2$$

$$P_n = P_0 \left(\frac{\rho_4^c}{\rho_3^c}\right)^{C_4} \left(\frac{\rho_3^c}{\rho_1^c}\right)^{C_3} \frac{\left(\rho_1^c\right)^n}{n!} \frac{1}{1-\lambda_4 E[B_4]} \frac{1}{1-\lambda_3 E[B_3]} \frac{1}{1-\lambda_2 E[B_2]} \qquad C_3 = C_2 \leqq n < C_1 = N$$

$$P_n = P_0 \left(\frac{\rho_4^c}{\rho_3^c}\right)^{C_4} \left(\frac{\rho_3^c}{\rho_1^c}\right)^{C_3} \frac{\left(\rho_1^c\right)^n}{n!} \frac{1}{1-\lambda_4 E[B_4]} \frac{1}{1-\lambda_3 E[B_3]} \frac{1}{1-\lambda_2 E[B_2]} \frac{1}{1-\lambda_1 E[B_1]} \qquad n = N = C_1$$

Finally, note that the stability conditions for the system are given by T equations of the form

$$\lambda_i < \left(E\left[R_{C_i}^{i-1} \right] \right)^{-1} = E[B_i]^{-1} \text{ for } 1 \leqq i \leqq T,$$

or, equivalently,

$$\rho_i < \mu \left(E\left[R_{C_i}^{i-1} \right] \right)^{-1} = \mu E[B_i]^{-1} \text{ for } 1 \leqq i \leqq T. \qquad (15)$$

The constraint on ρ_i depends both on the server cutoffs C_1 through C_i and on the partial loads ρ_j ($j<i$) of the higher priority arrival streams. If inequality (15) is violated for some i, then the system is unstable for all customers of priority k, where $k \geqq i$.

4.3 The Waiting Time Distributions

Consider an M/G/1 system with the following definitions:

$\tilde{S}(s)$ ≡ Laplace transform of the service time distribution

λ ≡ average customer arrival rate

E[B] ≡ average duration of a busy period

W ≡ waiting time in queue

It is well known (Conway [1967]; Kleinrock [1975], pp. 219-223) that the conditional Laplace-Stieltjes transform of in-queue waiting time W, conditioned on the arriving customer entering a busy period, is

$$E\left[e^{-sW} \,\middle|\, \textit{arrival during a busy period} \right] = \frac{1-\tilde{S}(s)}{\left(s-\lambda+\lambda\tilde{S}(s)\right)E[B]} \tag{16}$$

This result is directly applicable to the M/M/{C_i} system. For arrivals of priority i, let W_i denote the in-queue waiting time. In applying (16), λ should be replaced with λ_i and $\tilde{S}(s)$ with $\tilde{R}_{c_i}^{i-1}(s)=\tilde{B}_i(s)$. A busy period here, in M/G/1 parlance, corresponds to the continuous period of time during which the system is in superstate C_i^+, assuming $C_{i+1}<C_i$. If $C_{i+1}=C_i$, then the busy period for analysis of the priority i "M/G/1" queue is the "service time" of the priority i+1 "M/G/1" queue. In either case, the busy period for priority i, which we shall call the "level i" busy period, is the r.v. $R_{c_i}^{i}$. Thus, we can write

$$E\left[e^{-sW_i} \,\middle|\, \textit{arrival during a level i busy perio} \right] = \frac{1-\tilde{B}_i(s)}{\left(s-\lambda_i+\lambda_i\tilde{B}_i(s)\right)E\left[R_{C_i}^{i}\right]}$$

Finally, removing the condition that the arrival occurs during a level i busy period, using (8), and the fact that Poisson arrivals see time averages (Wolff [1982]),

we find:

$$\widetilde{W}_i(s) \equiv E\left[e^{-sW_i} \right] = \left(1 - P^+_{C_i} \right) + P^+_{C_i} \frac{1-\widetilde{B}_i(s)}{\left(s-\lambda_i+\lambda_i\widetilde{B}_i(s)\right)} \frac{1-\lambda_i E\left[B_i\right]}{E\left[B_i\right]} \tag{17}$$

For $i=1$, in particular, we can invert the transform $W_i(s)$ of the waiting time distribution, and we find, not surprisingly, that the waiting time is exponentially distributed; with a probability mass of $(1-P_N)$ at the origin,

$$\widetilde{W}_1(s) = (1-P_N) + P_N \frac{N\mu-\lambda_1}{N\mu-\lambda_1+s} \tag{18}$$

or,

$$W_1(y) \equiv P\{W_1 \leqq y\} = 1 - P_N e^{-(N\mu-\lambda_1)y} \qquad y \geqq 0 \tag{19}$$

It is, in general, difficult to invert the transforms and to obtain the waiting time distributions under closed form; we shall therefore content ourselves with first and second moments of the waiting times. As with M/G/1 systems, the k^{th} moment of the waiting time, W_i, is a function of the first $k+1$ moments of the service time, $R_{C_i}^{i-1} = B_i$.

Differentiating (17) and setting s to 0 (using l'Hopital's rule), we find the expected waiting time:

$$E\left[W_i\right] = P^+_{C_i} \frac{E\left[\left(B_i\right)^2\right]}{2E\left[B_i\right]} \frac{1}{1-\lambda_i E\left[B_i\right]} \tag{20}$$

Differentiating again, we obtain, after some algebra,

$$E\left[\left(W_i\right)^2\right] = 2E\left[W_i\right]^2 + \left(P^+_{C_i}\right)^2 \frac{E\left[\left(B_i\right)^3\right]}{3\lambda_i E\left[\left(B_i\right)^2\right]} \frac{1}{1-\lambda_i E\left[B_i\right]} \tag{21}$$

The first, second and third moments of B_i are derived in the appendix. For ease of reference, the recursions defining $E[B_i]$, $E[(B_i)^2]$ and $E[(B_i)^3]$ are also summarized there.

Our results for mean waiting times do not depend on the FIFO queue discipline within each priority class. Since the duration of a busy period is invariant under all workload conserving queueing disciplines, our equations for $E[W_i]$ are valid under any workload conserving queue disciplines for priorities 1,2,...,i-1 and, for priority i, under any workload conserving queue discipline that is independent of customer-specific service times. [Here we are referring to true service times, selected from the negative exponential distribution with mean $1/\mu$, not the B_i's.] As an example, suppose the priority 1 queue discipline is shortest job first (SJF), the priority 2 discipline is service in random order (SIRO), and the priority 3 is last come, first served (LCFS); then our equations for $E[W_2]$ and $E[W_3]$ remain valid, whereas our equation for $E[W_1]$ does not.

To conclude this section on waiting times, consider the special case $C_i = N\ \forall i$. The expected waiting time for priority i is given by (20). By repeated application of equations (8) and (27), we can write, for $j \in \{1,...,N\}$,

$$\frac{E\left[\left(R_N^j\right)^2\right]}{E\left[R_N^j\right]} = \frac{1}{\left(1-\lambda_j E\left[R_N^{j-1}\right]\right)^2} \frac{E\left[\left(R_N^{j-1}\right)^2\right]}{E\left[R_N^{j-1}\right]}$$

or,

$$\frac{E\left[\left(R_N^j\right)^2\right]}{E\left[R_N^j\right]} = \left(\prod_{k=1}^{j} \frac{1}{\left(1-\lambda_k E\left[R_N^{k-1}\right]\right)^2}\right) \cdot \frac{E\left[\left(R_N^0\right)^2\right]}{E\left[R_N^0\right]} \tag{22}$$

Therefore the expected in-queue waiting time for priority i customers can be written as

$$E\left[W_i\right] = \left(\prod_{k=1}^{i-1} \frac{1}{1-\lambda_k E\left[R_N^{k-1}\right]}\right)\left(\prod_{k=1}^{i} \frac{1}{1-\lambda_k E\left[R_N^{k-1}\right]}\right) \cdot \frac{P_N E\left[\left(R_N^0\right)^2\right]}{2E\left[R_N^0\right]} \tag{23}$$

Now, remarking that $f=\max\{k:C_k=N\}=N$ and $g=\min\{k:C_k=N\}=1$, we may apply equation (8) to equation (23), to obtain

$$E\left[W_i\right] = \frac{1}{1-\sum_{k=1}^{i-1}\lambda_k E\left[R_N^0\right]} \cdot \frac{1}{1-\sum_{k=1}^{i}\lambda_k E\left[R_N^0\right]} \cdot \frac{P_N E\left[\left(R_N^0\right)^2\right]}{2E\left[R_N^0\right]} \tag{24}$$

Under this form, when we replace $E[R_N^0]$ by $(N\mu)^{-1}$ and $E[(R_N^0)^2]$ by $2(N\mu)^{-2}$, our expected waiting time yields as a special case the famous Cobham formulas for prioritized M/M/N systems (Cobham [1954]),

$$E\left[W_i\right] = \frac{P_N/N\mu}{\left(1-\frac{1}{N\mu}\sum_{k=1}^{i-1}\lambda_k\right)\left(1-\frac{1}{N\mu}\sum_{k=1}^{i}\lambda_k\right)}$$

5 EXTENSION TO LOSS SYSTEMS

Armed with the methodology of Section 4, it is now easy to extend the results to M/M/{C_i} systems where customers of certain classes are queued if the system is busy (for their purposes) upon arrival, while customers of other arrival streams are lost under the same circumstances.

Essentially, all we need to do is to modify slightly our derivations of the distributions of the first passage times (the R_n^i's). Note that the recursions defining the respective Laplace-Stieltjes transforms, $R_n^i(s)$, are unchanged for $n \neq C_i$. We only need to concern ourselves with $R_{C_i}^i$, for $i=1,...,N$. If, by assumption, priority i customers are queued when they arrive while at least C_i servers are busy, then the results of Section 4 hold for $R_{C_i}^i$. If on the other hand they are lost, by assumption, then $R_{C_i}^i$ is the same as $R_{C_i}^{i-1}$. Therefore, in that case, equation (5) has to be modified to

$$\widetilde{R}_{C_i}^i(s) = \widetilde{R}_{C_i}^{i-1}(s) \qquad i=1,2,...,T \tag{25}$$

Consequently, equations (8), (32) and (35) must be replaced by, respectively,

$$E\left[R_{C_i}^i\right] = E\left[R_{C_i}^{i-1}\right] \qquad i=1,2,...,T \tag{26}$$

$$E\left[\left(R_{C_i}^i\right)^2\right] = E\left[\left(R_{C_i}^{i-1}\right)^2\right] \qquad i=1,2,...,T \tag{27}$$

$$E\left[\left(R_{C_i}^i\right)^3\right] = E\left[\left(R_{C_i}^{i-1}\right)^3\right] \qquad i=1,2,...,T \tag{28}$$

In order to obtain the steady state probabilities, P_n, of the number of busy servers, we apply equations (2) and (12) combined with equations (7), (9), and, depending on the queue/loss discipline, (8)/(26). It is easy to see that the P_n's are

then derived from:

$$P_n = P_{n-1}\frac{\rho_i^c}{n} \quad \text{for } 0 \leqq n < C_N \text{ and for } C_{i+1} \leqq n < C_i \text{ where } 1 \leqq i \leqq T-1 \tag{29}$$

$$P_n = P_{n-1}\frac{\rho_f^c}{n} \prod_{\substack{k:C_k = C_i \\ \text{priority } k \text{ queued}}} \left(\frac{1}{1-\lambda_k E\left[R_{C_i}^{k-1}\right]}\right) \quad \begin{array}{l}\text{for } n = C_i \text{ ,} \\ \text{where } f = \max\{k:C_k = C_i\}\end{array} \tag{30}$$

Consider, for example, a three priority system with $0<C_3<C_2<C_1=N$, where the priority 1 and the priority 3 customers queue (**Q**) if the system is busy when they arrive, while the priority 2 customers are lost (L) if the system is busy (i.e., at least C_2 servers are busy). Almost by inspection this QLQ-system has the following solution:

$$P_n = P_0 \frac{\left(\rho_3^c\right)^n}{n!} \qquad 0 \leqq n < C_3$$

$$P_n = P_0 \left(\frac{\rho_3^c}{\rho_2^c}\right)^{C_3} \frac{\left(\rho_2^c\right)^n}{n!} \frac{1}{1-\lambda_3 E\left[R_{C_3}^2\right]} \qquad C_3 \leqq n < C_2$$

$$P_n = P_0 \left(\frac{\rho_3^c}{\rho_2^c}\right)^{C_3}\left(\frac{\rho_2^c}{\rho_1^c}\right)^{C_2} \frac{\left(\rho_1^c\right)^n}{n!} \frac{1}{1-\lambda_3 E\left[R_{C_3}^2\right]} \qquad C_2 \leqq n < C_1$$

$$P_n = P_0 \left(\frac{\rho_3^c}{\rho_2^c}\right)^{C_3}\left(\frac{\rho_2^c}{\rho_1^c}\right)^{C_2} \frac{\left(\rho_1^c\right)^n}{n!} \frac{1}{1-\lambda_3 E\left[R_{C_3}^2\right]} \frac{1}{1-\lambda_1 E\left[R_N^0\right]} \qquad n = C_1 = N$$

where the $E[R_{C_i}^{i-1}]$ are obtained from the above equations, and P_0 is determined, as usually, by normalization.

We can derive moments of the waiting time distributions just as easily. Of course, for arrival streams that are lost if the system is busy, there is no waiting time; for queued arrival streams, on the other hand, equations (17) still holds. Therefore, the waiting times are computed in exactly the same way as in Section 4, but with the modified recursions (25), (26), (27), (28) for the first passage times R_n^i, where appropriate.

For the three-priority QLQ-system above, for example, the expected waiting times for priorities 1 and 3 are given by

$$E\left[W_i\right] = P_{C_i}^{+} \frac{E\left[\left(R_{C_i}^{i-1}\right)^2\right]}{2E\left[R_{C_i}^{i-1}\right]} \cdot \frac{1}{1-\lambda_i E\left[R_{C_i}^{i-1}\right]}$$

where the $E[R_n^i]$'s are derived from equations (7), (8)/(26) and (9).

6 COMPUTATIONAL RESULTS

We used the model to determine, for a hypothetical police department, the optimal cutoff number of patrol cars, C_i, beyond which no priority i calls should be served, lest the police response to higher priority calls be intolerably delayed. Our goal was to minimize a weighted sum of the expected waiting times for the various arrival streams, heavily weighting delays incurred by higher priority arrivals:

$$Minimize\, Z_1 = \sum_{i=1}^{T} K_i \frac{\lambda_i}{\lambda} E\left[W_i\right] \quad ,$$

$$where\, \lambda = \sum_{i=1}^{T} \lambda_i = \lambda_1^c \quad and \quad K_1 \geqq K_2 \geqq \ldots \geqq K_T \geqq 0.$$

(This objective function does not, of course, make sense for loss systems, e.g., the QLQ-system of Section 5.) Alternatively, one can, for example, minimize a weighted sum of the probabilities that a priority i arrival encounters a busy system:

$$Minimize\, Z_2 = \sum_{i=1}^{T} D_i \frac{\lambda_i}{\lambda} P_{C_i}^{+} \quad ,where\, D_1 \geqq D_2 \geqq \ldots \geqq D_T \geqq 0 \ .$$

Since we essentially optimize whatever objective function we choose by implicit enumeration of the solutions of all feasible M/M/{C_i} systems, objectives other than the above can be equally easily implemented. For certain objective functions, such as expected weighted waiting times, monotonicity and convexity arguments can substantially reduce the number of cases to be enumerated. We have not, however, investigated this objective-specific branch and bound procedure in any detail.

The computational results below were obtained on a PR1ME 850 computer. The largest problem we solved was a 25 server problem with 5 priorities. The run computed statistics and performance measures for some 2070 M/M/{C_i} queueing systems in 76 seconds of CPU time. The high speed of execution, we felt, did not warrant taking advantage of objective-specific improvements in our enumeration scheme.

Table 1 shows computational results for a three-priority system with nine servers. The arrival streams are characterized by $\lambda_1/\mu = 3$, $\lambda_2/\mu = 1$ and $\lambda_3/\mu = 2$. A maximum number of 9 servers are available, but results are also shown for a total of 8, 7 and 6 servers available. (In terms of patrol units this determines the optimal cutoffs for the case when one or more units are not operational for one reason or another.) We assume infinite queue capacity, so that all calls are queued if they find the system busy. For illustrative purposes we use the objective or cost functions

$$Z_1 = 100 \frac{\lambda_1}{\lambda} E[W_1] + 10 \frac{\lambda_2}{\lambda} E[W_2] + \frac{\lambda_3}{\lambda} E[W_3] \quad , and$$

$$Z_2 = 25 \frac{\lambda_1}{\lambda} P_N + 5 \frac{\lambda_2}{\lambda} P_{C_2}^{+} + \frac{\lambda_3}{\lambda} P_{C_3}^{+} \quad .$$

Table 1 summarizes the most important system statistics, as well as the above "performance measures." The optimal values for Z_1 and Z_2 respectively are marked by an asterisk.

C_1	C_2	C_3	$P^+_{C_1}$	$P^+_{C_2}$	$P^+_{C_3}$	$E[W_1]$	$E[W_2]$	$E[W_3]$	Z_1	Z_2	
9	9	9	0.1960	0.1960	0.1960	0.0326	0.0587	0.1176	3.541	5.357	
9	9	8	0.1388	0.1388	0.3123	0.0231	0.0416	0.2186	2.598	3.910	
9	9	7	0.1022	0.1022	0.4855	0.0170	0.0306	0.4820	2.127	3.050	
9	9	6	0.0779	0.0779	0.7115	0.0130	0.0233	1.363	2.286	2.554	
9	9	5	0.0617	0.0617	0.9685	0.0102	0.0185	18.08	13.14	2.292	
9	8	8	0.1075	0.3224	0.3224	0.0179	0.1075	0.2457	2.313	3.439	
9	8	7	0.0787	0.2362	0.4922	0.0131	0.0787	0.5152	1.918*	2.691	total of 9 servers
9	8	6	0.0599	0.1799	0.7158	0.0100	0.0600	1.422	2.147	2.276	
9	8	5	0.0474	0.1423	0.9711	0.0079	0.0474	19.95	14.25	2.071*	
9	7	7	0.0621	0.5178	0.5178	0.0104	0.2138	0.6472	2.180	2.762	
9	7	6	0.0470	0.3915	0.7321	0.0078	0.1616	1.659	2.428	2.315	
9	7	5	0.0370	0.3089	0.9808	0.0062	0.1275	31.63	22.13	2.095	
9	6	6	0.0378	0.7846	0.7846	0.0063	0.4424	2.678	3.889	2.775	
8	8	8	0.3570	0.3570	0.3570	0.0714	0.1428	0.3570	7.854	9.757	
8	8	7	0.2572	0.2572	0.5145	0.0515	0.1029	0.6431	5.917	7.203	
8	8	6	0.1947	0.1947	0.7299	0.0389	0.0779	1.642	5.248	5.677	total of 8 servers
8	7	7	0.2011	0.5362	0.5362	0.0402	0.2423	0.7709	5.343	6.278	
8	7	6	0.1512	0.4033	0.7436	0.0303	0.1822	1.867	4.877*	4.949	
8	6	6	0.1212	0.7945	0.7945	0.0242	0.4671	2.972	5.963	4.884*	
7	7	7	0.6138	0.6138	0.6138	0.1535	0.3581	1.432	17.49	16.78	
7	7	6	0.4520	0.4520	0.7910	0.1130	0.2637	2.966	14.16	12.58	total of 7 servers
7	6	6	0.3577	0.8346	0.8346	0.0894	0.5646	4.517	13.84*	10.89*	

Table 1

Statistics and Performance Measures for a System With Up to Nine Servers ($\frac{\lambda_1}{\mu} = 3$, $\frac{\lambda_2}{\mu} = 1$, $\frac{\lambda_3}{\mu} = 2$).

APPENDIX

Moments of First Passage Times

In section 4, we derived the Laplace transforms for what we loosely call the "first passage times" R_n (equations (4), (5) and (6):

$$\widetilde{R}^{o}_{N}(s) = \frac{N\mu}{N\mu+s} \tag{4}$$

$$\widetilde{R}^{i}_{C_i}(s) = R^{i-1}_{C_i}\left(s+\lambda_i-\lambda_i\widetilde{R}^{i}_{C_i}(s)\right) \qquad i=1,2,\ldots,T \tag{5}$$

$$\widetilde{R}^{i}_{n}(s) = n\mu\left[\lambda^{c}_{i}+n\mu+s-\lambda^{c}_{i}\widetilde{R}^{i}_{n+1}(s)\right]^{-1} \qquad \begin{matrix} C_{i+1}\le n<C_i \\ 1\le i<T \end{matrix} \tag{6}$$

Remember also that by definition

$$\widetilde{B}_i(s)=\widetilde{R}^{i-1}_{C_i}(s) \qquad i=1,2,\ldots,T \ .$$

The steady state probabilities and the moments of the waiting times for the various priorities derived in Section 4 are expressed in terms of the moments of the R_n 's. Their first, second and third moments are derived here. The algebra is very straightforward, albeit rather unedifying. For notational convenience define

$$\alpha_i(s)\equiv s+\lambda_i-\lambda_i\widetilde{R}^{i}_{C_i}(s) \quad \textit{and} \quad \beta^{i}_{n}(s)\equiv\lambda^{c}_{i}+n\mu+s-\lambda^{c}_{i}\widetilde{R}^{i}_{n+1}(s) \ .$$

Differentiating equations (5) and (6) once with respect to s yields

$$\frac{d\widetilde{R}^{i}_{C_i}(s)}{ds} = \left(1-\lambda_i\frac{d\widetilde{R}^{i}_{C_i}(s)}{ds}\right)\frac{d\widetilde{R}^{i-1}_{C_i}(\alpha_i(s))}{d(\alpha_i(s))} \ ,$$

and,

$$\frac{d\widetilde{R}^{i}_{n}(s)}{ds} = -n\mu\left(1-\lambda^{c}_{i}\frac{d\widetilde{R}^{i}_{n+1}(s)}{ds}\right)\beta^{i}_{n}(s)^{-2} \ .$$

Setting s to zero, the first moments of the recursions are given by

$$E\left[R_N^0 \right] = \frac{1}{N\mu} \tag{7}$$

$$E\left[R_{C_i}^i\right] = \frac{E\left[R_{C_i}^{i-1}\right]}{1-\lambda_i E\left[R_{C_i}^{i-1}\right]} \qquad i=1,2...T \tag{8}$$

$$E\left[R_n^i\right] = \frac{1}{n\mu}\left(1+\lambda_i^c E\left[R_{n+1}^i\right]\right) \qquad \begin{matrix} i=1,2,...,T-1 \\ C_{i+1}\le n<C_i \end{matrix} \tag{9}$$

Differentiating equations (5) and (6) a second time yields:

$$\frac{d^2\widetilde{R}_{C_i}^i(s)}{ds^2} = -\lambda_i \frac{d^2\widetilde{R}_{C_i}^i(s)}{ds^2}\frac{d\widetilde{R}_{C_i}^{i-1}(\alpha_i(s))}{d(\alpha_i(s))} + \left(1-\lambda_i\frac{d\widetilde{R}_{C_i}^i(s)}{ds}\right)^2 \frac{d^2\widetilde{R}_{C_i}^{i-1}(\alpha_i(s))}{d(\alpha_i(s))^2} \quad ,$$

and,

$$\frac{d^2\widetilde{R}_n^i(s)}{ds^2} = n\mu\lambda_i^c \frac{d^2\widetilde{R}_{n+1}^i(s)}{ds^2}\beta_n^i(s)^{-2} + 2n\mu\left(1-\lambda_i^c\frac{d\widetilde{R}_{n+1}^i(s)}{ds}\right)^2\beta_n^i(s)^{-3} \quad .$$

Setting s to zero, we find the second moments:

$$E\left[\left(R_N^0\right)^2\right] = \frac{2}{(N\mu)^2} \quad , \tag{31}$$

$$E\left[\left(R_{C_i}^i\right)^2\right] = \frac{\left(1+\lambda_i E\left[R_{C_i}^i\right]\right)^2}{1-\lambda_i E\left[R_{C_i}^{i-1}\right]} E\left[\left(R_{C_i}^{i-1}\right)^2\right] \qquad i=1,2...T \tag{32}$$

$$E\left[\left(R_n^i\right)^2\right] = \frac{\lambda_i^c}{n\mu}E\left[\left(R_{n+1}^i\right)^2\right] + 2E\left[R_n^i\right]^2 \qquad \begin{matrix} i=1,2,...,T-1 \\ C_{i+1}\le n<C_i \end{matrix} \tag{33}$$

Finally, differentiating equations (5) and (6) a third time yields:

$$\frac{d^3\widetilde{R}^i_{C_i}(s)}{ds^3} = -\lambda_i \frac{d^3\widetilde{R}^i_{C_i}(s)}{ds^3}\frac{d\widetilde{R}^{i-1}_{C_i}(a_i(s))}{d(a_i(s))} - 3\lambda_i \frac{d^2\widetilde{R}^i_{C_i}(s)}{ds^2}\left(1-\lambda_i\frac{d\widetilde{R}^i_{C_i}(s)}{ds}\right)\frac{d^2\widetilde{R}^{i-1}_{C_i}(a_i(s))}{d(a_i(s))^2}$$

$$+\left(1-\lambda_i\frac{d\widetilde{R}^i_{C_i}(s)}{ds}\right)^3\frac{d^3\widetilde{R}^{i-1}_{C_i}(a_i(s))}{d(a_i(s))^3} \qquad i=1,2,\ldots,T$$

and,

$$\frac{d^3\widetilde{R}^i_n(s)}{ds^3} = n\mu\lambda_i^c\frac{d^3\widetilde{R}^i_{n+1}(s)}{ds^3}\beta^i_n(s)^{-2} - 6n\mu\lambda_i^c\frac{d^2\widetilde{R}^i_{n+1}(s)}{ds^2}\left(1-\lambda_i^c\frac{d\widetilde{R}^i_{n+1}(s)}{ds}\right)\beta^i_n(s)^{-3}$$

$$-6n\mu\left(1-\lambda_i^c\frac{d\widetilde{R}^i_{n+1}(s)}{ds}\right)^3\frac{d^3\widetilde{R}^i_{n+1}(s)}{ds^3}\beta^i_n(s)^{-4} \qquad \begin{matrix} i=1,2,\ldots,T-1 \\ C_{i+1}\le n<C_i \end{matrix}.$$

Whence the third moments of the first passage times, setting s to zero,

$$E\left[\left(R^0_N\right)^3\right] = \frac{6}{(N\mu)^3} \quad , \tag{34}$$

$$E\left[\left(R^i_{C_i}\right)^3\right] = \frac{\left(1+\lambda_i E\left[R^i_{C_i}\right]\right)^3}{1-\lambda_i E\left[R^{i-1}_{C_i}\right]}E\left[\left(R^{i-1}_{C_i}\right)^3\right] + 3\lambda_i\frac{1+\lambda_i E\left[R^i_{C_i}\right]}{1-\lambda_i E\left[R^{i-1}_{C_i}\right]}E\left[\left(R^i_{C_i}\right)^2\right]E\left[\left(R^{i-1}_{C_i}\right)\right] \quad 1\le i<T \tag{35}$$

$$E\left[\left(R^i_n\right)^3\right] = \frac{\lambda_i^c}{n\mu}E\left[\left(R^i_{n+1}\right)^3\right] + 6E\left[\left(R^i_n\right)^2\right]E\left[R^i_n\right] - 6E\left[R^i_n\right]^3 \qquad \begin{matrix} i=1,2,\ldots,T-1 \\ C_{i+1}\le n<C_i \end{matrix} \tag{36}$$

ACKNOWLEDGMENT

This work was supported in part by the National Institute of Justice, U.S. Department of Justice (Grant No. 83-IJ-CX-0065) and in part by the National Science Foundation, (Grant No. 8411871-SES).

REFERENCES

ABOL'NIKOV, L.M., R.M. YASNOGORODSKIY. 1974. A Class of Queueing Problems with Priorities When There are Urgent Orders. *Eng. Cyb.* 12, 62-72.

BENN, B.A. Hierarchical Car Pool Systems in Railroad Transportation. Ph.D. thesis, Case Institute of Technology, Cleveland, OH. 1966.

COBHAM, A. 1954. Priority Assignment In Waiting Line Problems. *Opns. Res.* 2, 70-76.

CONWAY, R.W., W.L. MAXWELL AND L.W. MILLER. 1967. *Theory of Scheduling.* Addison-Wesley, Reading, MA.

COOPER, R.B. 1972. *Introduction to Queueing Theory.* Macmillan, New York, NY.

ESOGBUE, A.O. AND A.J. SINGH. 1976. A Stochastic Model for an Optimal Priority Bed Distribution Problem in a Hospital Ward. *Opns. Res.* 24, 885-898.

JAISWAL, N.K. 1968. *Priority Queues.* Acad. Press, New York, NY.

KLEINROCK, L. 1975. *Queueing Systems, Vol. 1 and 2.* John Wiley and Sons, Inc., New York, NY.

LARSON, R.C. 1972. *Urban Police Patrol Analysis.* MIT Press, Cambridge, MA.

MCCLAIN, J.O. 1976. Bed Planning Using Queueing Theory Models of Hospital Occupancy: A Sensitivity Analysis. *Inquiry.* 13, 167-176.

SHONICK, W. AND J.R. JACKSON. 1973. An Improved Stochastic Model for Occupancy-Related Random Variables in General-Acute Hospitals. *Opns. Res.* 21, 952-965.

TAYLOR, I.D.S. AND J.G.C. TEMPLETON. 1980. Waiting Time In a Multi-Server Cutoff-Priority Queue, and Its Application to an Urban Ambulance Service. *Opns. Res.* 28, 1168-1188.

TIEN, J.M. et al. 1976. An Evaluation Report: Wilmington Split Force Patrol Program. Public Systems Evaluation, Inc., Cambridge, MA.

WOLFF, R.W. 1982. Poisson Arrivals See Time Averages. *Opns. Res.* 30, 223-231.