THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 10, December 1975
Printed in U.S.A.

An Approximate Method for Calculating Delays for a Family of Cyclic-Type Queues

By S. HALFIN

(Manuscript received May 22, 1975)

A study of the marker-register dial-tone delay problem in No. 5 crossbar switching machines led to a special type of cyclic queuinq model. In this paper, we present a method for calculating approximately the steady-state delays of an arriving customer. When applied to the marker-register problem, the model emphasizes the order in which markers are assigned to waiting calls and the fact that part of the markers' time is unproductive when an "all registers busy" condition occurs. Some numerical results are presented, which agree with the observed phenomenon that, for a constant marker load, the delays of Touch-Tone® *calls are influenced by the load on the dial-pulse originating registers, and vice versa. The results are compared to those of a simulation of the same problem. The numerical results compare favorably in the range of loads that produce a dial-tone speed of between 0.05 and 0.15.*

I. INTRODUCTION

The queuing model described in this paper resulted from a study of the marker-register dial-tone delay problem in No. 5 crossbar switching machines. A number of queues with Poisson arrivals of equal rates are served in a cyclic order by a server with constant service time. Upon arriving at a nonempty queue, the server chooses a customer from the queue at random. After one service time, the customer either leaves the system with a certain predetermined probability or rejoins his queue. In both cases, the server uses a fixed amount of time and moves to the next queue; thus, at most one customer leaves the system following each arrival of the server at a queue.

Related models were treated by Cooper,[1] Cooper and Murray,[2] and Eisenberg.[3] In Refs. 1 and 2, the server either empties the queue being served or serves all those present at the queue in its arrival epoch. The case of two queues with different arrival rates is treated in Ref. 3. In these papers, the Laplace-Stieltjes transforms of the waiting time distributions were obtained. Attempts to obtain the distributions for similar models by approximate methods were made by Leibowitz[4] and Schay, Jr.[5]

The method presented here approximately calculates the steady-state delays of an arriving customer. The approximation is carried out by modifying the model to achieve a manageable state space. Next, the method is applied to the problem of dial-tone delay in the No. 5 crossbar switching machine. The order in which markers are assigned to waiting calls and the fact that part of the markers' time is unproductive when an "all registers busy" condition occurs are emphasized in the model. The numerical results presented agree with the observed phenomenon that, for a constant marker load, the delays of *Touch-Tone®* calls are influenced by the load on the dial-pulse originating registers, and vice versa. The results are compared to those of a simulation of the same problem.

Several features of the No. 5 crossbar machine, which may have an influence on the dial-tone speed, were excluded. Some of these features were investigated in subsequent work by H. A. Guess[6] and are described in more detail in Section XVI of this paper.

II. THE MODEL

Let N queues $E_1, E_2, \cdots, E_N$, $N \geqq 2$, be given. Customer arrivals to each queue constitute independent Poisson processes, all having the same rate λ. The queues are served by a single server in the following way: At each point in time, the server is at some queue. Transitions in its position occur at discrete time epochs which are equally spaced with periods of duration T. At such time epochs, the server moves instantaneously to the next nonempty queue in a circular order, chooses a waiting customer in that queue at random, and stays with this customer until the end of the period. The served customer then leaves the queue with probability p^*, or remains in the queue with probability $q^* = 1 - p^*$. If all the queues are empty at a transition point, then no change occurs in the position of the server.

We assume further that T is small with respect to the accuracy with which we want to know the delays, and thus all the arrivals can be assumed to occur at the transition epochs and the queuing process can be considered in discrete time. Thus, if $X_{i,k}$ is the number of arrivals to E_i at the kth time epoch, then all $X_{i,k}$, $1 \leqq i \leqq N, k \geqq 0$, are independent identically distributed random variables, all having a Poisson distribution with mean λT.

III. THE FULL STATE SPACE

The state of the queuing system at any time epoch is defined as the $(N+1)$tuple $(m_1, m_2, \cdots, m_N, n)$, where m_i is the number of waiting customers at E_i, and n is the position of the server *before* the transition. We call the set of all such states the full state space. It is then clear from the discussion in the previous section that our queuing system, with the full state space, is a stationary Markov chain. Thus, in principle, one can calculate transient and steady-state probabilities. However, the full state space is much too large for practical computational purposes.

Consider, for instance, the dial-tone delay problem where a typical number of N is 15. Then, even if the system is so underloaded that each m_i can be restricted to be either 0 or 1, we have $2^{15} \times 15 \sim 500{,}000$ states. A natural approach, which we follow in the remainder of the paper, is to approximate the behavior of our system with systems having a smaller size state space.

IV. THE "BLACK BOX" APPROACH

Some important facts about the system can be deduced by considering only the total number of customers in all the queues $s = m_1 + m_2 + \cdots + m_N$. It is clear that the system with this single state is again a stationary Markov chain. It is, in fact, a discrete version of an $M/D/1$ queue, with the added feature that a customer who is held by the server returns to the queue with probability q^*. Alternatively, the service time measured in units of T may be considered as having a geometrical distribution with mean $1/p^*$. The traffic intensity of the system is the $\rho = \lambda NT/p^*$; thus, we have Theorem 1.

Theorem 1: A necessary and sufficient condition for the nonsaturation of the system is $\lambda NT < p^*$.

Note that, because of the symmetry, a particular queue in the system is saturated if and only if the system as a whole is saturated; hence, Theorem 1 provides a saturation condition for all the individual queues.

Let us denote by A_i^* the probability that a Poisson-distributed random variable with mean λNT attains the value i. The system has the following transition probabilities. For $s > 0$,

$$\Pr\,(s \to s') = A^*_{s'-s+1}p^* + A^*_{s'-s}q^*$$

and

$$\Pr\,(0 \to s') = A_{s'}.$$

The equations for the steady-state probabilities P_s are then

$$P_{s'} = A^*_{s'}P_0 + \sum_{s=1}^{s'+1} (A^*_{s'-s+1}p^* + A^*_{s'-s}q^*)P_s \qquad s' = 0, 1, \cdots. \quad (1)$$

Equations (1) can be solved recursively, starting with $P_0 = 1 - \rho$.

One can also calculate the generating function

$$\hat{P}(u) = \sum_{s=0}^{\infty} P_s u^s$$

$$= \frac{(p^* - \alpha)(1-u)}{uq^* + p^* - ue^{\alpha(1-u)}},$$

where $\alpha = \lambda NT$. By evaluating $\hat{P}'(1)$, the expected value of the total number of customers in the system $\bar{s}$ is

$$\bar{s} = \frac{2\alpha - \alpha^2}{2(p^* - \alpha)}. \tag{2}$$

V. THE MODIFIED MODEL

We now consider our original system with a new state space. A state will now consist of a triplet (m, M, n), where $m = m_1$, $M = m_2 + m_3 + \cdots + m_N$, and n is the same as previously, namely the position of the server before a transition. It is clear that the new state space is much smaller than the full state space; however, the Markovian property is lost. This can be seen by the following argument: If M was positive at time $k-1$, and if in the transition between $k-1$ to k the server skipped a large number of queues, then those M customers were concentrated in the remaining queues; thus, it is probable that, in the k to $k+1$ transition, a small number of queues will be skipped.

At this point, we modify our model to make it a Markov chain with respect to the new state space. To do this, we need to define one-step transition probabilities so that the behavior of the modified model will approximate that of the original model. Let the position of a customer be the queue number where he waits. Our key assumption concerns the probability distribution of the positions of the M customers, given the state (m, M, n).

For the remainder of this section, let us enumerate the queues $E_2, \cdots, E_N$ by starting with the queue following the position of the server and observing the cyclic order, skipping E_1. Next, we make the following assumptions: Let $M > 0$; then

(*i*) The positions of the M customers are independent, identically distributed, random variables.

(*ii*) The probability that any one of the M customers will be in the ith queue (in the new order) is $\pi_i(M)$, where

$$\pi_i(M) = b(M) + \frac{(N-i-1)R(M)}{M(N-2)} \qquad i = 1, 2, \cdots, N-1,$$

where $b(M) = 1/(N-1) - [R(M)/2M]$ and determination of $R(M)$ is described below.

The rationale behind assumption (*ii*) is that a customer is less likely to be in a queue that has just recently been visited. The average difference between the number of customers in the first and last queues is, according to assumption (*ii*), $M[\pi_1(M) - \pi_{N-1}(M)] = R(M)$, which should be approximately equal to the expected number of arrivals during one full cycle of the server. Thus,

$$R(M) \sim \lambda T R_0(M),$$

where $R_0(M)$ is the expected number of nonempty queues, given M. Finally, we approximate $R_0(M)$ by

$$R_0(M) = (N-1)\left[1 - \left(\frac{N-2}{N-1}\right)^M\right],$$

where the right-hand side is the expected number of nonempty queues among $E_2, \cdots, E_N$, if the M customers are uniformly distributed. We conclude this section by using the new assumptions to calculate some probabilities that will be needed later.

Let J denote the number of successive empty queues following the position of the server (when E_1 is disregarded), $J = 0, 1, \cdots, N-1$.

The distribution of J depends on M. Let

$$\begin{aligned} Q_j(M) &= \Pr\,(J \geqq j-1) \\ q_j(M) &= \Pr\,(J = j-1) \\ j &= 1, 2, \cdots, N. \end{aligned}$$

Using assumptions (*i*) and (*ii*), we have

$$Q_j(M) = \left[\sum_{i=j}^{N-1} \pi_i(M)\right]^M \qquad j = 1, \cdots, N-1$$

$$Q_N(M) = \begin{cases} 0 & \text{if} \quad M > 0 \\ 1 & \text{if} \quad M = 0, \end{cases}$$

and

$$\begin{aligned} q_j(M) &= Q_j(M) - Q_{j+1}(M) \qquad j = 1, \cdots, N-1 \\ q_N(M) &= Q_N(M). \end{aligned}$$

VI. TRANSITION PROBABILITIES FOR THE MODIFIED SYSTEM

Given a state (m, M, n), the transition probabilities of the position of the server can be expressed in terms of the $Q_j(M)$ and $q_j(M)$'s as follows.

For $m > 0$,

$$\Pr(n \to n') = \begin{cases} q_{n'-n}(M) & \text{if} \quad n' > n \\ Q_{N-n+1}(M) & \text{if} \quad n' = 1 \\ 0 & \text{if} \quad 1 < n' \leqq n. \end{cases}$$

For $m = 0$, $M > 0$,

$$\Pr(n \to n') = \begin{cases} q_{n'-n}(M) & \text{if} \quad n' > n \\ 0 & \text{if} \quad n' = 1 \\ q_{N+n'-n-1} & \text{if} \quad 1 < n' \leqq n, \end{cases} \tag{3}$$

and, for $m = 0$, $M = 0$,

$$\Pr(n \to n') = \begin{cases} 1 & \text{if} \quad n' = n \\ 0 & \text{if} \quad n' \neq n. \end{cases}$$

Let us denote by A_i and a_i the probability that Poisson-distributed random variables with means $\lambda(N-1)T$ and λT, respectively, attain the value i. We can now write the state transition probabilities.

For $m > 0$, $M > 0$:

$$\Pr[(m, M, n) \to (m', M', n')] = \begin{cases} a_{m'-m} q_{n'-n}(M)[A_{M'-M+1}p^* + A_{M'-M}q^*] & \text{if} \quad n' > n \\ A_{M'-M} Q_{N+1-n}(M)[a_{m'-m+1}p^* + a_{m'-m}q^*] & \text{if} \quad n' = 1, \\ 0 & \text{otherwise.} \end{cases}$$

For $m = 0$, $M > 0$:

$$\Pr[(0, M, n) \to (m', M', n')] = a_{m'}[A_{M'-M+1}p^* + A_{M'-M}q^*] \times \begin{cases} q_{n'-n}(M) & \text{if} \quad n' > n \\ 0 & \text{if} \quad n' = 1 \\ q_{N+n'-n-1}(M) & \text{if} \quad 1 < n' \leqq n. \end{cases}$$

For $m > 0$, $M = 0$:

$$\Pr[(m, 0, n) \to (m', M', n')] = \begin{cases} A_{M'}[a_{m'-m+1}p^* + a_{m'-m}q^*] & \text{if} \quad n' = 1, \\ 0 & \text{if} \quad n' \neq 1, \end{cases}$$

and finally, for $m = M = 0$,

$$\Pr[(0, 0, n) \to (m', M', n')] = \begin{cases} a_{m'} A_{M'} & \text{if} \quad n' = n \\ 0 & \text{if} \quad n' \neq n. \end{cases} \tag{4}$$

VII. STEADY-STATE EQUATIONS FOR THE MODIFIED SYSTEM

If we consider the total number of customers in the system $s = m + M$, then it is clear that all the results of Section IV remain valid for the modified system. In particular, $\lambda NT < p^*$ is the necessary and sufficient condition for the existence of a steady state. In the following discussion, we assume that this condition is satisfied. Let $P(m, M, n)$ be the steady-state probability of the state (m, M, n) satisfying the following equations.

For $n' > 1,\ m' \geqq 0,\ M' \geqq 0$:

$$P(m', M', n')$$

$$= \sum_{n=1}^{n'-1} \sum_{m=0}^{m'} \sum_{M=1}^{M'+1} a_{m'-m} q_{n'-n}(M)[A_{M'-M+1}p^* + A_{M'-M}q^*]$$

$$\times P(m, M, n) + \sum_{n=n'}^{N} \sum_{M=1}^{M'+1} a_{m'}[A_{M'-M+1}p^* + A_{M'-M}q^*]$$

$$\times q_{N+n'-n-1}(M) P(0, M, n) + a_{m'} A_{M'} P(0, 0, n'). \quad (5)$$

For $n' = 1,\ m' \geqq 0,\ M' \geqq 0$:

$$P(m', M', 1) = \sum_{n=1}^{N} \sum_{m=1}^{m'+1} \sum_{M=0}^{M'} A_{M'-M} Q_{N+1-n}(M)[a_{m'-m+1}p^* + a_{m'-m}q^*]$$

$$\times P(m, M, n) + a_{m'} A_{M'} P(0, 0, 1).$$

VIII. NUMERICAL SOLUTION OF THE STEADY-STATE EQUATIONS

Equations (5) can be solved either as a system of linear equations with the auxiliary equation

$$\sum_{m,M,n} P(m, M, n) = 1$$

or by starting with any initial distribution and iterating it through (4) until a desired degree of convergence is obtained.

It seems adequate to adopt the second method since the steady state of the modified system is of interest to us only as an approximation to the steady state of the original system. Thus, an extensive computational effort to obtain an accurate solution to (4) is not warranted. A good initial distribution to start the iterations can be obtained in the following way. We have

$$\sum_{m+M=s} \sum_{n=1}^{N} P(m, M, n) = P_s \qquad s = 0, 1, \cdots,$$

where the P_s were computed by the method outlined in Section IV. If we divide the s customers uniformly among the queues and make the position of the server random, we get the following distribution:

$$P_0(m, s-m, n) = \frac{1}{N} B\left(m, \frac{1}{n}, s\right) P_s$$

$$s = 0, 1, \cdots \qquad m = 0, \cdots, s \qquad n = 1, \cdots, N,$$

where $B(m, 1/N, s)$ is the binomial probability of m successes out of s trials, with probability of success $= 1/N$.

IX. EVALUATION OF THE DELAYS FOR THE MODIFIED SYSTEM

Let $f(m, M, n, k)$ be the probability that a customer arriving at E_1 will wait k service periods before leaving the system, given that on his arrival the system went into the state (m, M, n). The customer stays for at least one service period, so $k \geqq 1$. Notice also that $m \geqq 1$.

Theorem 2: The delay probabilities satisfy the following recursive equations:

$$f(m, M, n, 1) = \frac{p^*}{m} Q_{N-n+1}(M)$$

and

$$f(m, M, n, k+1) = \sigma_1 + \sigma_2$$

$$m \geqq 1, \quad M \geqq 0, \quad 1 \leqq n \leqq N; \qquad k \geqq 1,$$

where

$$\sigma_1 = \begin{cases} \sum_{n'=n+1}^{n} \sum_{M'=M-1}^{\infty} \sum_{m'=m}^{\infty} a_{m'-m} q_{n'-n}(M) [A_{M'-M+1} p^* + A_{M'-M} q^*] \\ \qquad \times f(m', M', n', k) \quad \text{if} \quad M > 0,\ n < N \\ 0 \quad \text{if} \quad M = 0 \text{ or } n = N \end{cases}$$

$$\sigma_2 = \sum_{M'=M}^{\infty} \sum_{m'=m-1}^{\infty} A_{M'-M} Q_{N+1-n}(M) \left[\frac{m-1}{m} a_{m'-m+1} p^* + a_{m'-m} q^* \right] \times f(m', M', 1, k).$$

Proof: $f(m, M, n, 1)$ equals the probability that the server moves to E_1, that the particular customer is selected for service, and that he leaves the system after the service period. Hence, the formula for $k = 1$.

For $k \geqq 1$, we have

$$f(m, M, n, k+1) = \sum_{m', M', n'} \Pr [(m, M, n) \to (m', M', n') \cap \text{the customer stays in } E_1] f(m', M', n', k).$$

Using eq. (4), we get that σ_1 is the part of the right-hand side corresponding to $n' \neq 1$ and σ_2 is the part of the right-hand side corresponding to $n' = 1$.

Theorem 2 provides a method for calculating the delays conditional on the state. Let $f^*(k)$ be the probability that a customer arriving at E_1 will wait k service periods before leaving the system, given that before his arrival the system was in the steady state.

Theorem 3:

$$f^*(k) = \sum_{m'=1}^{\infty} \sum_{M'=0}^{\infty} \sum_{n'=1}^{N} f(m', M', n', k) P^*(m', M', n') \qquad k = 1, 2, \cdots,$$

where

$$P^*(m', M', n') = \frac{1}{1 - a_0} \bigg[P(m', M', n') - a_0 \sum_{n=1}^{n'-1} \sum_{M=1}^{M'+1} q_{n'-n}(M) \cdot (A_{M'-M+1} p^* + A_{M'-M} q^*) P(m', M, n) \bigg]$$

and

$$P^*(m', M', 1) = \frac{1}{1 - a_0} \bigg[P(m', M', 1) - a_0 \sum_{n=1}^{N} \sum_{M=0}^{M'} Q_{N+1-n}(M) \cdot [p^* P(m' + 1, M, n) + q^* P(m', M, n)] \bigg]$$

for $m' \geqq 1$, $M' \geqq 0$, $2 \leqq n' \leqq N$.

Proof: The theorem is valid if it is shown that $P^*(m', M', n')$ is the probability of the state (m', M', n') at the point of arrival of the customer, say, point u, for all $m' \geqq 1$, $M' \geqq 0$, $1 \leqq n' \leqq N$. The probabilities of states with $m' = 0$ is zero, since there is at least one customer in E_1. We know that at $u - 1$ the system was in a steady state. The transition probabilities between $u - 1$ and u, conditional that at least one customer arrives at E_1, are obtained from (4) by replacing a_i with a_i^*, where

$$a_i^* = \begin{cases} \dfrac{1}{1 - a_0} a_i & i = 1, 2, \cdots \\ 0 & i = 0. \end{cases}$$

The expressions for $P^*(m', M', n')$ can now be calculated by operating on the steady-state probabilities with the modified transition probabilities and using the fact that the steady-state probabilities satisfy eqs. (5).

X. VALIDITY OF THE MODIFIED SYSTEM

The difference between the original and the modified systems is in the rules of the server movement. In the modified system, the server does not follow the cyclic order. However, to calculate the delay distribution, we are interested only in the pattern of the time points when the server is in E_1, and that, hopefully, is similar to the corresponding pattern in the original system. The degree of similarity is difficult to check, except by simulating the original system. Verification of the "reasonability" of the modified system may be made by checking whether each E_n receives the same number of visits by the server, and if E_1 gets the right amount of expected number of waiting customers, i.e.,

$$\sum_{m=0}^{\infty} \sum_{M=0}^{\infty} P(m, M, n) \sim \frac{1}{N} \qquad n = 1, \cdots, N$$

and

$$\sum_{m=0}^{\infty} m \sum_{M=0}^{\infty} \sum_{n=1}^{N} P(m, M, n) \sim \frac{\bar{s}}{N},$$

where $\bar{s}$ is given by eq. (2).

XI. THE MARKER-REGISTER SYSTEM

In the rest of the paper, we apply the model to the dial-tone delay problem in the No. 5 crossbar switching machine. Following are some operational features of the No. 5 crossbar switching machines which are relevant to the dial-tone delay distribution.

Calls appear on line link frames (LLF). The dial tone markers (DTM) which are not busy are paired to the waiting calls. Under "normal" operation, i.e., when several DTMs are free, the LLFs look for available DTMs according to a fixed preference order. When all the DTMs become busy, a gate is closed and the DTMs serve first those LLFs that contain waiting calls at that moment. If an LLF has more than one call waiting, only one call will be served during the gating period.

When a DTM becomes idle following the "all markers busy" condition, it looks for a waiting call according to the following scheme. Each DTM has its own order in which it scans the LLFs. Thus, for example, when there are four DTMs and 60 LLFs, the first DTM will scan the LLFs in the natural order from 0 to 59, the second DTM will start at LLF 15, go to 59, and then come back to 0 to 14, etc.*

When a DTM locates an LLF with waiting calls, it chooses one of those calls and proceeds to look for an originating register (OR) for the call. The above choice may be considered random for all practical purposes.† If the DTM finds a vacant OR, then it connects the call to the OR, and the calling customer gets a dial tone. If no OR is available, then the DTM releases the call, and it continues to wait in its LLF and to bid for a DTM. In both cases, the holding time of the DTM is constant and approximately equal. We denote this time by T, where T is approximately 0.25 second. In fact, this time is approximately 0.21 second in

* This is the recommended arrangement, although not all No. 5 crossbar entities observe it.

† We omit consideration of the systematic preference for serving calls in vertical group 2.

the case where no OR can be found, but we ignore this difference to simplify our model.

We also assume that the distribution of holding times of the ORs is negative exponential for a conservative estimate of the delay distribution. The arrival of calls to each LLF is assumed to be Poisson, with the rate being equal for all LLFs. We denote the rate for a single LLF by λ. Finally, in many cases there are two types of calls, dial-pulse and *Touch-Tone*, where both types are served by the same DTMs but require different ORs. When there are two types of calls, the ratio between their arrival rates is assumed to be the same in every LLF.

XII. A QUEUING MODEL FOR ONE TYPE OF CALL

The system described in the previous section is quite complicated, and it appears that, to model such a system and be able to derive numerical results from the model, some simplifying assumptions are inevitable. One such model was proposed by W. S. Hayward.[7] Its basic assumption is that, in order to be served, a call must find both a marker and a register idle. Once the marker and register start processing a call, they act independently of each other, each having exponentially distributed holding times. To solve the resulting state equations, Hayward introduced a system with one type of server, which approximates the behavior of his model.

The present queuing model emphasizes the order in which the markers are assigned to waiting calls, and takes into account the fact that the time a marker spends serving a call is nearly the same, whether or not it found a free register.

First we assume that each DTM serves only those LLFs which are of high priority on its list. Thus, in the example of the previous section, the first DTM will serve only the first 15 LLFs, the second DTM will serve only the next 15 LLFs, etc. Such an assumption is justified under heavy load conditions. We denote by N the number of LLFs which are served by one DTM.

Next, we assume that each DTM serves its LLFs in a cyclic order and that, whenever it finds a LLF with waiting calls, it serves exactly one call. This assumption is asymptotically valid under heavy traffic loads, because of the gating procedure described in the previous section.

Finally, we assume that whenever a DTM serves a call there is a fixed probability p^* that an OR will be available and thus that the waiting time of the call will end (i.e., the customer gets a dial tone). This assumption would hold if the availability of the ORs is independent of the number of waiting calls, which is clearly not the case. This assumption will cause our model to somewhat underestimate the delays, while the first assumption tends to overestimate them. The value of p^* can be taken approximately as the delay probability of a call, given that the arrival process of the calls to the ORs is Poisson, and therefore it can be computed by the Erlang C formula.

Thus, we arrive at the model which was described in Section II, with the server the DTM. The server chooses a customer (call) from the queue at random. After one service time, the customer either leaves the system with probability p^* or rejoins his queue (that is, waits in his LLF). The server then moves to the next LLF having a waiting call.

XIII. SOME THEORETICAL AND NUMERICAL RESULTS

It was shown in Section IV that the occupancy of the DTM is

$$\rho = \frac{\lambda NT}{p^*}.$$

Hence, a necessary and sufficient condition for nonsaturation of the system is $\lambda NT < p^*$. Also, the expected total number of waiting calls in the LLFs which are served by the DTM is

$$\bar{s} = \frac{\rho}{2(1-\rho)}(2 - \rho p^*).$$

Thus, for a fixed occupancy of the DTM, the expected total number of waiting calls is a monotone decreasing function of p^*. The same is true for the expected delay, $\bar{W}$, since by Little's formula

$$\bar{W} = \frac{\bar{s}}{\lambda TN} = \frac{\bar{s}}{\rho p^*},$$

and so

$$\bar{W} = \frac{1}{2(1-\rho)}\left(\frac{2}{p^*} - \rho\right).$$

A standard measure for the quality of service is the dial-tone speed (DTS), which is the probability that the call will have to wait three seconds or more for a dial tone. Figure 1 presents some computed values of the DTS for various values of p^* and λ with ρ held constant at three different values. It is seen that DTS is also a monotone decreasing function of p^*, for a fixed DTM occupancy.

XIV. A MODEL FOR TWO TYPES OF CALLS

Consider now the case of a system having dial-pulse and *Touch-Tone* calls; this is the usual situation in No. 5 crossbar offices today. Let the arrival rates from each LLF be λ_1 and λ_2, and let the probabilities of finding available registers be p_1^* and p_2^* for dial-pulse and *Touch-Tone*

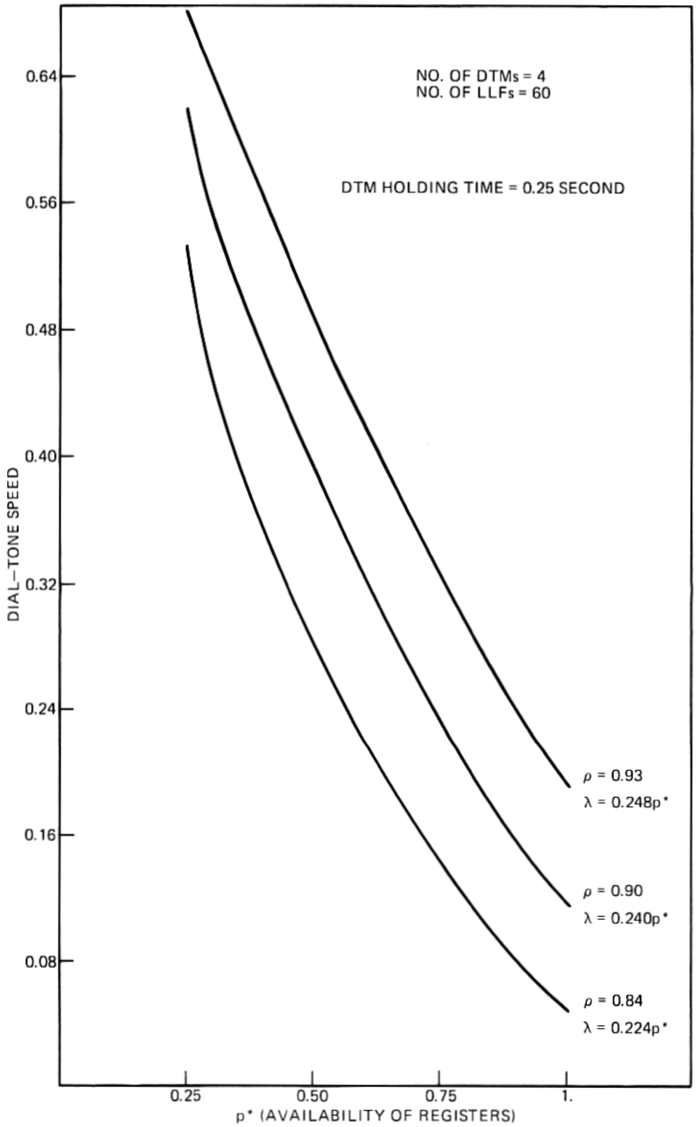

Fig. 1—DTS as a function of p^* and λ for constant DTM occupancy ρ (queuing model).

calls, respectively. The loads on the DTM due to the two types of calls are

$$\rho_1 = \frac{\lambda_1 NT}{p_1^*} \qquad \text{and} \qquad \rho_2 = \frac{\lambda_2 NT}{p_2^*}.$$

To approximate the delays of this system, we consider a system with one type of call, for which $\lambda = \lambda_1 + \lambda_2$ and $\rho = \rho_1 + \rho_2$. The appropriate p^* of this system satisfies

$$p^* = \frac{\lambda_1 + \lambda_2}{\lambda_1/p_1^* + \lambda_2/p_2^*}.$$

Now the state probabilities can be computed by the approximate method; however, it is necessary to modify the formulas for the computation of the delay distributions to obtain those distributions conditional on the type of call.

Let $f_i(m, M, n, k)$ be the probability that a call of type i ($i = 1, 2$) arriving at LLF No. 1 will wait k service periods before leaving the system, given that on its arrival the system went into the state (m, M, n). The recursive formulas of Theorem 2 have to be modified to

$$f_i(m, M, n, 1) = \frac{p_i^*}{m} Q_{N-n+1}(M)$$

and

$$f_i(m, M, n, k + 1) = \sigma_1 + \sigma_2$$
$$m \geqq 1, \quad M \geqq 0, \quad 1 \leqq n \leqq N, \quad k \geqq 1,$$

where σ_1 is as in Section IX except that f is replaced by f_i and

$$\sigma_2 = \sum_{m'=M}^{\infty} \sum_{m'=m-1}^{\infty} A_{M'-M} Q_{N+1-n}(M) \left[\frac{m-1}{m} a_{m'-m+1} p^* + a_{m'-m}\left(\frac{m-1}{m} q^* + \frac{q_i^*}{m}\right)\right] [f_i(m', M', 1, k)].$$

The proof of the validity of the modified formulas is along the same lines as the proof of Theorem 2.

XV. NUMERICAL RESULTS

Several computer runs were made for a typical large system with 60 LLFs, 4 DTMs ($N = 15$), 100 dial-pulse ORs and 50 *Touch-Tone* ORs, with both dial-pulse ORs and *Touch-Tone* ORs having a mean holding time of 13 seconds. T was taken to be 0.25 second in all runs. The parameters varied were $\lambda = \lambda_1 + \lambda_2$, the total input rate per LLF, and $\alpha = \lambda_1/\lambda_2$ (the ratio of the rates of the two types of calls).

Figures 2 and 3 describe the results for $\alpha = 2$, that is, when the ratio of the loads is the same as that of the ORs. Figure 2 describes the behavior of the occupancies of the ORs and the DTM, while in Fig. 3 the DTS is plotted as a function of λ. Figures 4 and 5 present the corresponding results for $\alpha = 3$. The values of λ were chosen to be near the point of saturation, i.e., where the occupancy of the DTM ap-

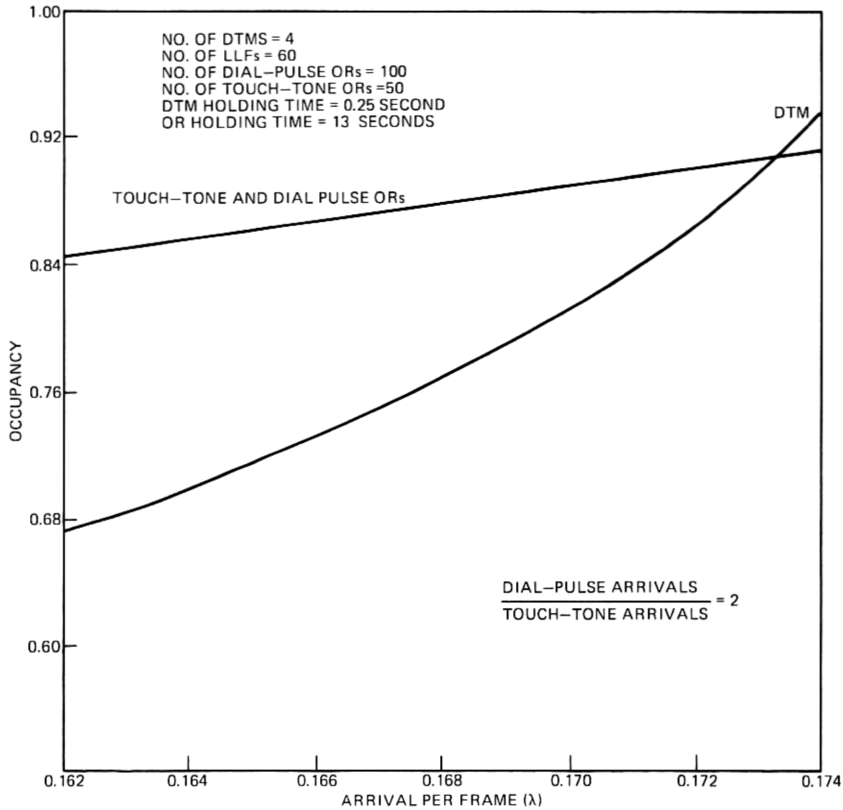

Fig. 2—Occupancies of the DTM and ORs as a function of λ (queuing model).

proaches 1. In this range, the DTS is sensitive to small perturbations in λ.

Figure 6 shows the dependence of the DTS on α for a fixed λ ($\lambda = 0.158$ per second). It can be seen that the quality of service deteriorates as α diverges from the neighborhood of the ratio of the number of dial-pulse ORs/number of *Touch-Tone* ORs (which equals 2 in our case). This is consistent with the observation that *Touch-Tone* delays are significantly influenced by the dial-pulse OR load for a constant offered load to the DTM.

The results were compared with those of a simulation model which we constructed for the system described in Section XI. Tables I and II present a comparison between the results of the queuing model and the simulation. We compare the intensities of input for which levels of the DTS are reached between 0.05 and 0.25 for a balanced system and between 0.05 and 0.15 for an unbalanced system. Examining those

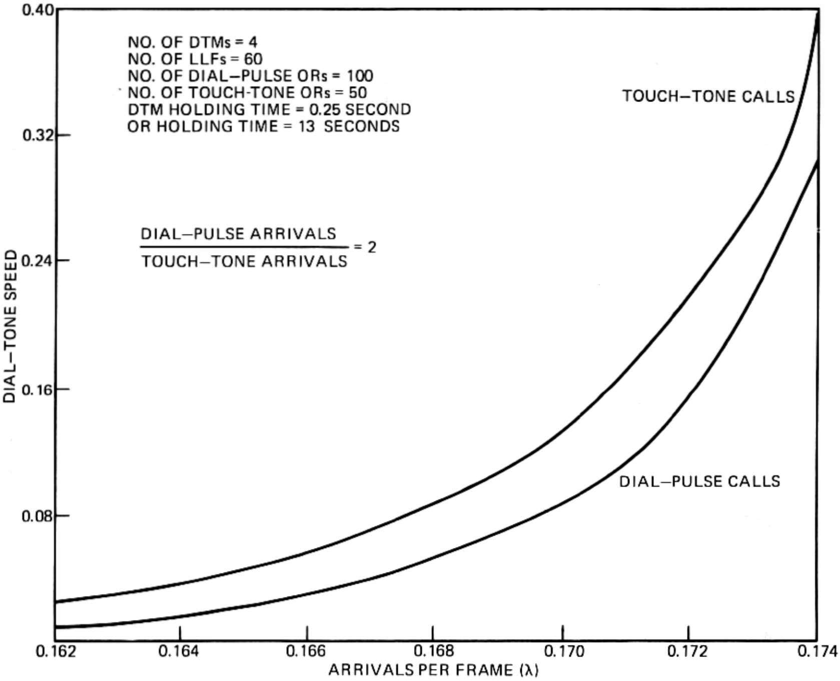

Fig. 3—DTS as functions of λ (queuing model).

Table I — Values of λ for which given levels of DTS are reached, for $\lambda_1/\lambda_2 = 2$

(a) Dial-Pulse Calls

DTS	Queuing Model	Simulation
0.05	0.1680	0.1695
0.10	0.1705	0.1735
0.15	0.1720	0.1760
0.20	0.1730	0.1775
0.25	0.1735	0.1790

(b) *Touch-Tone* Calls

DTS	Queuing Model	Simulation
0.10	0.1690	0.1690
0.15	0.1705	0.1720
0.20	0.1720	0.1740
0.25	0.1725	0.1760

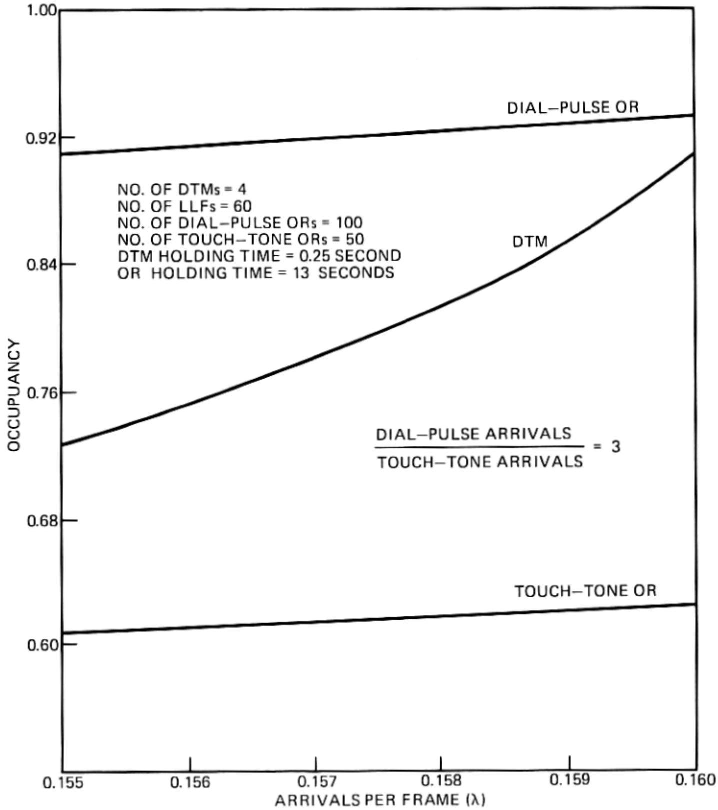

Fig. 4—Occupancies of the DTM and ORs as a function of λ (queuing model).

Table II — Values of λ for which given levels of DTS are reached, for $\lambda_1/\lambda_2 = 3$

(a) Dial-Pulse Calls

DTS	Queuing Model	Simulation
0.05	0.1540	0.1550
0.10	0.1565	0.1590
0.15	0.1580	0.1610
0.20	0.1585	0.1620

(b) *Touch-Tone* Calls

DTS	Queuing Model	Simulation
0.05	0.1575	0.1600
0.10	0.1590	0.1625
0.15	0.1600	0.1640

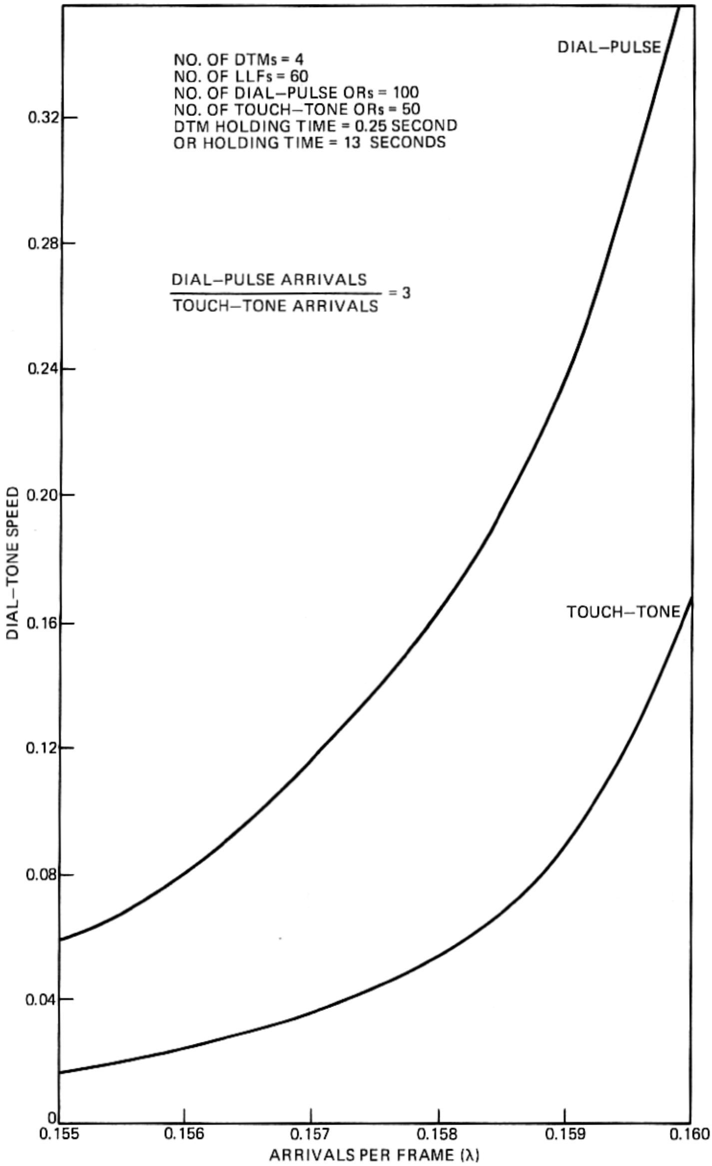

Fig. 5—DTS as functions of λ (queuing model).

tables, we observe that the differences in the corresponding figures for the two models in these regions are less than 4 percent of the total input. Also, it can be observed that the computed DTS grows faster in the queuing model than in the simulation. The reason is that the

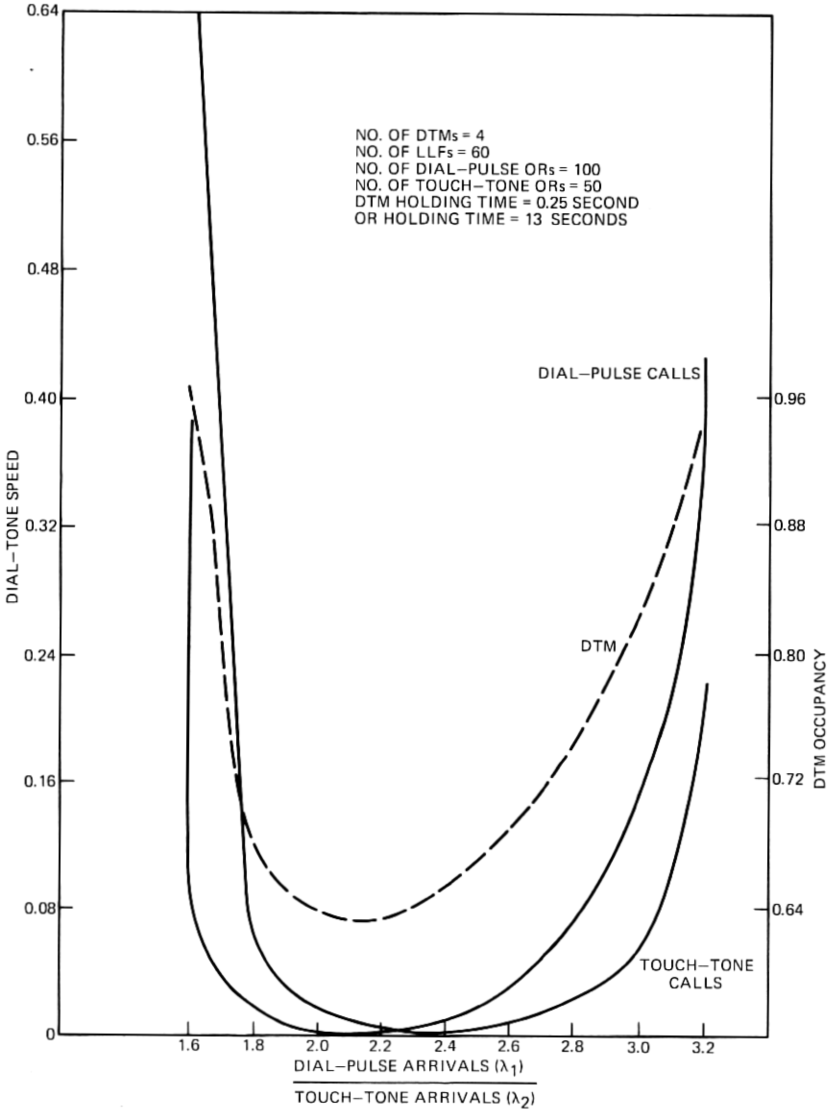

Fig. 6—DTS and DTM occupancy as functions of λ_1/λ_2 for a fixed λ, $\lambda = 0.158$.

assumption made in Section XII, that the probability of finding all registers busy can be computed by the Erlang C formula, is incorrect in this region. Figure 7 compares the DTS as computed by the two models. Again we conclude that the fit is fair in the "critical" region. It would have been desirable to validate the results by performing a field trial. Such a trial should consist of measuring the DTS for a No. 5

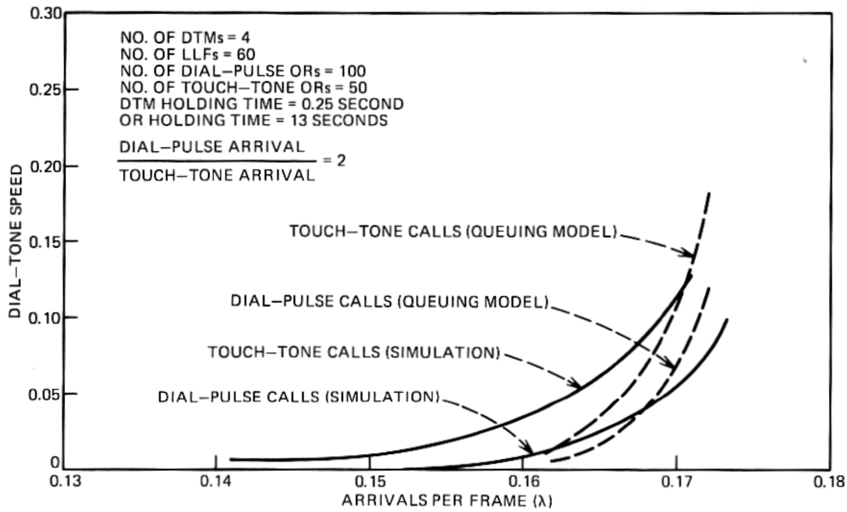

Fig. 7—DTS as function of λ (queuing model and simulation).

crossbar machine with high constant loads. However, observing Fig. 8, where the cumulative distributions of the hourly DTS from the simulation were plotted, one sees that these distributions have long tails. This implies that, to get a reasonably accurate estimate for the DTS, say, with a standard error of 1 percent, we would have to run the trials for around 25 hours while keeping the load constant. This seems to be a difficult task.

XVI. SUMMARY AND CONCLUSIONS

We presented a method for modifying the original model, as presented in Section II, to a model which has a much smaller state space. Methods were described for calculating the steady-state distributions of the states and of the delays in the modified model. The model was applied to the problem of calculating dial-tone delays in the No. 5 crossbar switching machine. This was accomplished by making some simplifying assumptions about the order of service of the waiting calls by the markers. The numerical results were compared to those of a simulation, and found to be close on an important range of the DTS. This gives us a certain amount of confidence that both models are valid, which is especially important because of the difficulty in validating the models by experimental data, as discussed in Section XV. However, the reader should be aware that several features of the No. 5 crossbar machine, which may have an influence on the DTS, were excluded.

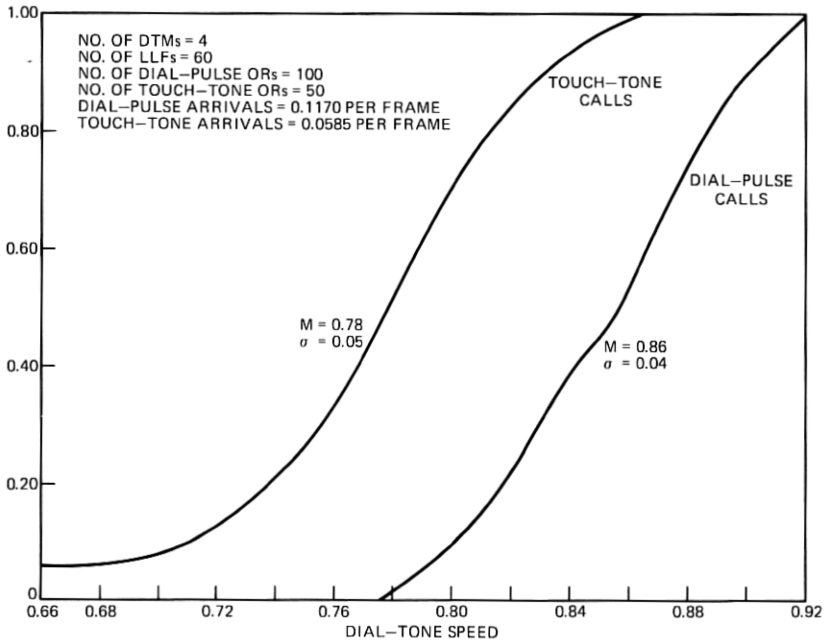

Fig. 8—Cumulative distributions of the hourly dial-tone speed (simulation).

A major effect not covered by the present model is the effect of horizontal group blocking on dial-tone speed and on dial-tone marker waste usage. Recently obtained field data and theoretical studies reported in a subsequent paper by H. A. Guess[6] have shown that dial-tone speed and dial-tone marker occupancy can be appreciably increased by horizontal group blocking caused by high average line link frame loads and also by poor load balance. Consequently, the dial-tone speeds associated with a given call origination rate in an actual No. 5 crossbar office may be higher than would be predicted by our model.

XVII. ACKNOWLEDGMENTS

I wish to thank J. G. Kappel and M. F. Morse for introducing me to the dial-tone delay problem.

REFERENCES

1. R. B. Cooper, "Queues Served in Cyclic Order: Waiting Times," B.S.T.J., *49*, No. 3 (March 1970), pp. 399–414.
2. R. B. Cooper and G. Murray, "Queues Served in Cyclic Order," B.S.T.J., *48*, No. 3 (March 1969), pp. 675–689.
3. M. Eisenberg, "Two Queues with Alternating Service," unpublished work.

4. M. A. Leibowitz, "An Approximate Method for Treating a Class of Multiqueue Problems," IBM Journal, *5*, July 1961, pp. 204–209.
5. G. Schay, Jr., "Approximate Methods for a Multiqueuing Problem," IBM Journal, *6*, April 1962, pp. 246–249.
6. H. A. Guess, "The Effect of Frame Load and Balance on Dial-Tone Delay in No. 5 Crossbar," B.S.T.J., this issue, pp. 1755–1793.
7. W. S. Hayward, "Capacities of Interacting Server Groups in Common Control Switching Systems," Proceedings of the Sixth International Teletraffic Congress, Munich, September 9–15, 1970, pp. 325/1–6.