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

The Effect of Frame Load and Balance on Dial-Tone Delay in No. 5 Crossbar

By H. A. GUESS

(Manuscript received May 22, 1975)

Recently obtained field data and theoretical studies show that, for a fixed subscriber calling rate, dial-tone delay in No. 5 crossbar can be appreciably increased by high average line-link frame loads and also by poor load balance. The increased delay is caused by waste dial-tone-marker usage generated by a small number of calls that encounter horizontal group blocking in obtaining a dialing connection. This paper discusses an analytical model to predict the time from receiver off-hook to receipt of dial tone under various service conditions in No. 5 crossbar.

I. INTRODUCTION

1.1 Description of the dial-tone connection process

In a No. 5 crossbar switching machine, dial tone is provided to a subscriber line, terminating on a line-link frame (LLF), by an originating register (OR), terminating on a trunk-link frame (TLF), via a series of three network links: line links, junctors, and trunk links. The dialing connections are set up by dial-tone markers (DTMs), which are common control devices. Each line-link frame contains a number of crossbar switches that are used to establish connections between subscriber lines and trunks, or between subscriber lines and service circuits, such as originating registers. The crossbar switches that form line concentrators on which groups of subscriber lines terminate are called horizontal groups. Maximum size offices typically contain from 4 to 6 DTMs, 40 to 60 LLFs, 20 to 30 TLFs, and up to 140 ORs. Each LLF contains 10 horizontal groups and each horizontal group is a concentrator containing between 29 and 59 subscriber lines on the input side of the switch and 10 line links on the output side of the switch.

To provide dial tone to a subscriber, the off-hook signal from the subscriber line initiates a bid for a DTM through a connector circuit unique to each LLF. As soon as a DTM becomes available, it locates an unoccupied OR and then attempts to find a dialing path (consisting of a line link, a junctor, and a trunk link) connecting the OR with the subscriber line. As soon as the connection is established, the DTM releases and proceeds to serve other calls waiting for dial tone. The OR provides dial tone, receives the dialed digits, obtains a completing marker, transmits information to it, and releases. The completing marker then establishes the connection between the calling subscriber and an outgoing or intraoffice trunk.

If all ORs are busy, the DTM releases and the call rejoins what is effectively a queue of calls waiting for a DTM. If an OR is available, but no (unoccupied) dialing path connecting the OR with the subscriber line can be found, the DTM will release the OR, obtain a second (usually different) OR, and try to find a dialing path between that OR and the subscriber line. When a DTM is unable to find a dialing path between a given OR and a given subscriber line, a matching failure is said to occur. If, on the second try, the DTM cannot find a dialing path, the DTM releases and the call rejoins the queue of calls waiting for a DTM. In such a case, a DTM second-failure-to-match (DTM2FTM) is said to occur.

The method of assigning DTMs to waiting calls is controlled by a type of "gating" circuitry designed to equalize service and to reduce the incidence of long delays in obtaining dial tone. When one or more DTMs are free (light traffic operation), the LLFs look for DTMs according to a fixed preference order. When all DTMs become busy and a request for a DTM occurs, a gate is closed and only the LLFs containing calls waiting for dial tone at that moment are put inside the gate. The DTMs then proceed to serve the LLFs inside the gate. If more than one call is waiting on an LLF, only one call will be served during the gating period. Once a call on an LLF is served during a given gating period, that LLF is put out of the gate, whether or not the DTM is successful in establishing a dialing connection for the call. When all LLFs with requests at the start of the gating cycle have been put out of the gate, the gate opens; if there are sufficient waiting calls to cause all of the DTMs to become busy again, a new gating cycle will be started; otherwise, light traffic operation will resume. During a gating period, a DTM that becomes idle scans the LLFs in cyclic order. Each DTM has a different starting LLF for the scan so as to equalize service. (The description of the gating procedure is taken from Refs. 1 and 2.)

1.2 Effects of matching failures on dial-tone delay

Repetitive matching failures can occur in establishing the dialing connection essentially because the first-stage crossbar switch on which the subscriber line terminates (the horizontal group) is a concentrator whose output links (line links) have holding times that are much longer than the holding times of the common control devices that set up the dialing connections (the DTMs). The average line-link holding times, being largely determined by conversation holding times (with allowance for ineffective attempts), can be on the order of 150 seconds or more, while the DTM holding times are typically on the order of 0.25–0.40 second.

Although matching failures (blocking) can occur when one or more of the 10 line links on a horizontal group are unoccupied, *repetitive* matching failures under such conditions are rare because a DTM is quite likely to be successful in establishing a dialing connection after a few attempts. The expected holding time in a blocked condition of a call that finds all 10 line links busy is the lesser of the length of time for one of the 10 line links to become free (with a short added time for DTM uses and matching failures in setting up the dialing path) and the length of time that a subscriber is willing to wait.

Since, at average busy-hour load levels, less than about one percent of all originating calls are predicted to encounter an all-10-line-links-busy condition, previous dial-tone-delay studies have assumed the effect of matching failures on dial-tone delay to be small. However, the fact that line-link holding times are much longer than DTM holding times means that a call which finds all 10 line links busy can remain blocked for long enough to consume a large number of DTM uses (except at calling rates sufficiently high that very little of the offered blocked-call load is carried by the DTMs). Thus, it is possible for a small number of calls experiencing blocking to generate a disproportionate number of waste DTM uses, increase DTM occupancy, and thereby increase dial-tone delay for all other calls in the office.

II. ANALYTICAL MODEL (LIMITING FORM)

2.1 Assumptions

Since the main effect of matching failures on dial-tone delay is caused by the resulting waste DTM use, and since the dial-tone delay distribution of the small proportion of calls experiencing repetitive matching failures can be calculated approximately (see Section 5.2), the effect of matching failures on the dial-tone delay distribution of calls that do not experience repetitive matching failures has been represented in terms of a queuing model with two classes of calls, good calls and bad calls, defined as follows:

(*i*) A *good call* experiences no matching failures but is subject to delay caused by DTMs and ORs.

(*ii*) A *bad call* experiences total network blocking (no dialing connection available) and defects from the system after an exponential waiting time.

We will first describe the mathematical structure of the analytical model used for dial-tone delay calculations. We will refer to this model as *the limiting model.* Next, we derive the DTM saturation load* and prove that it is not changed by the presence of bad calls.

In Appendix B, we prove that the equilibrium queue length and waiting time distributions of the limiting model are the limits in distribution of a sequence of equilibrium distributions arising from a model in which the expected bad-call arrival rate approaches zero and the expected bad-call waiting time (until defection) approaches infinity in such a way that their product, total erlangs of bad calls, is constant.

In the limiting model, a good call arrives and finds a random (truncated) Poisson-distributed (but time-independent) number of bad calls permanently present in the system.† The queue discipline is characterized by random order of service. The DTMs cannot distinguish between good and bad calls when choosing a call to be served. This corresponds to the fact that, in the No. 5 crossbar switching machine, a DTM cannot recognize that a 10-line-links-busy condition exists on a particular horizontal group and, hence, cannot avoid wasting time serving calls for which no dialing path exists. Equilibrium good-call queue length distributions are computed conditional upon the number of bad calls present in the system. Calculating the expectation of the conditional distribution over the distribution of bad calls gives the unconditional dial-tone-delay distribution for good calls. The conditional distributions depend on the total office calling rate and on the number of LLFs, DTMs, and ORs but do not depend on the horizontal group load or the load variation. Hence, delay distributions for a range of frame load and balance effects can be calculated using the same set of conditional distributions, thereby greatly reducing the computer time needed for parametric studies.

Since DTM holding times are approximately constant (for a given set of office parameters and traffic characteristics) and since these holding times are small with respect to the accuracy with which it is necessary to be able to predict delays, we treat the dial-tone delay process as a discrete time queue with a constant service time of T seconds. Good calls are assumed to arrive in batches according to a Poisson process at times kT, for $k = 1, 2, \cdots$. Immediately upon arrival, the good calls join the queue of good and bad calls waiting for dial tone. Calls are chosen at random from the queue for service

* The DTM saturation load is defined to be the good-call originating load below which a steady state good-call-queue-length distribution exists and above which such a steady state distribution does not exist.

† When the bad-call input is assumed to be peaked, the number of bad calls in the system has a (truncated) negative binomial distribution. This is discussed in Section IV.

by the DTMs with each call—whether a good call or a bad call—having an equal probability of being chosen. A good call served on the DTM cycle beginning at time kT will either acquire an OR, receive dial tone, and thereupon exit the system at time $(k + 1)T$, or else will fail to acquire an OR and will return to the queue of calls waiting for dial tone at time $(k + 1)T$. A bad call served on the DTM cycle beginning at time kT will return to the queue of calls waiting for dial tone at time $(k + 1)T$. Note that since new arrivals occur only at the time points kT, a DTM that is idle on the cycle beginning at time kT will remain idle at least until the start of the cycle beginning at time $(k + 1)T$. The dial-tone delay for a call is the length of time from the moment when the call arrives to when it obtains dial tone and (simultaneously) leaves the dial-tone queue. Thus, the minimum possible dial-tone delay in the model is T seconds.

In the analytical model, two DTMs serving N line-link frames are used to represent four DTMs serving $2N$ line-link frames. A single DTM holding time equal to the office average DTM holding time is used for both good and bad calls. In actuality and in the simulation models, bad calls have somewhat longer DTM holding times than good calls and good calls that encounter a condition of all-ORs-busy have somewhat shorter DTM holding times than good calls that do not encounter a condition of all-ORs-busy. Comparison of results from the analytical model with those of the gating simulation model indicates that these simplifications tend to offset each other in the range of interest.

A further simplification in the analytical model concerns the manner in which availability of ORs is treated. Since No. 5 crossbar offices typically contain over 100 ORs and since the ORs are in tandem with the DTMs, it is presently not practical to keep track of the number of occupied ORs directly in an analytical model. Availability of ORs is treated by assuming that at each time point kT, all ORs are busy with probability q and two or more ORs are free with probability $p = 1 - q$. The calculated probability that exactly one OR is free and that a dialing connection is available between this one OR and the given subscriber line is sufficiently small, in the occupancy range of interest, so as not to warrant the additional complexity caused by introducing this effect.

The idea of using a discrete time model and the method of treating OR availability through use of a fixed probability of all ORs busy are due to Halfin.[3] Following Halfin, the probability q is taken to be the erlang C probability of all ORs busy. At OR occupancies above about 0.90 with frame loads low enough that few second failures to match occur, this method of treating OR availability somewhat underpredicts the delay caused by an all-registers-busy condition (based on comparison with simulation results). The underprediction arises because, at higher OR occupancies, once all ORs become busy they tend to remain busy for a time period equal to several DTM cycles, as one would expect from the erlang *C* delay formula.

Preliminary studies were made using an analytical single-server cyclic queuing model, developed by S. Halfin,[3] which represented the No. 5 crossbar gating process in considerable detail but did not take into account the effects of horizontal group blocking. These studies showed that the delays predicted by the cyclic queuing model do not differ appreciably from those of a discrete-time M/D/1 queue with feedback and random order of service. The latter model requires about 1/30th of the computer time required by the former. Both models overpredicted simulations. For these reasons, explicit representation of the gating process was not attempted and a discrete-time M/D/2 queue with feedback and random order of service was taken as the starting point for developing an analytical dial-tone-delay model to include effects of horizontal group blocking.

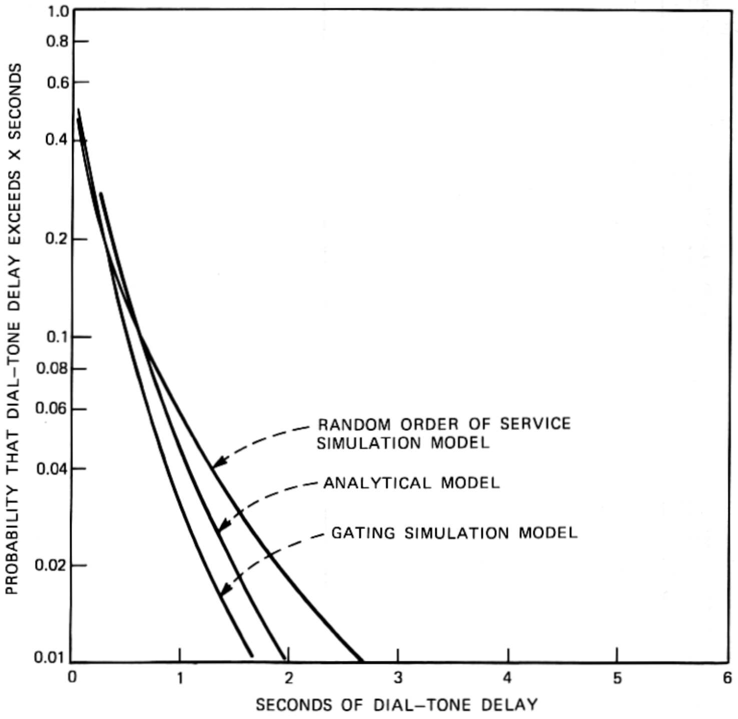

Fig. 1—Comparison of analytical and simulation models. Dial-tone delays are at 1400 CCS/LLF and 0.70 DTM OCC (excluding DTM2FTMS).

Comparison of simulation results with predictions using the limiting analytical model (based on two markers and random order of service) shows that the predicted delays typically fall somewhat above or very close to simulated delays based on four markers with gating order of service and somewhat below simulated delays based on four markers with random order of service. In the light of these results and because of the large scatter in actual measured No. 5 crossbar dial-tone delays, it did not seem worth the considerable added complexity to include explicit representation of the gating process in the analytical model. Typical results are shown in Figs. 1 and 2.

2.2 Queue length equations for the limiting model

The queue-length process for good calls in the limiting model is a discrete-time Markov chain with finitely many irreducible classes $\{C_k\}$ that are noncommunicating in the sense that no transition between different classes is possible. Hence, each class is itself an irreducible Markov chain. The kth class is the queue-length Markov chain for a discrete-time M/D/2 queue with random order of service,

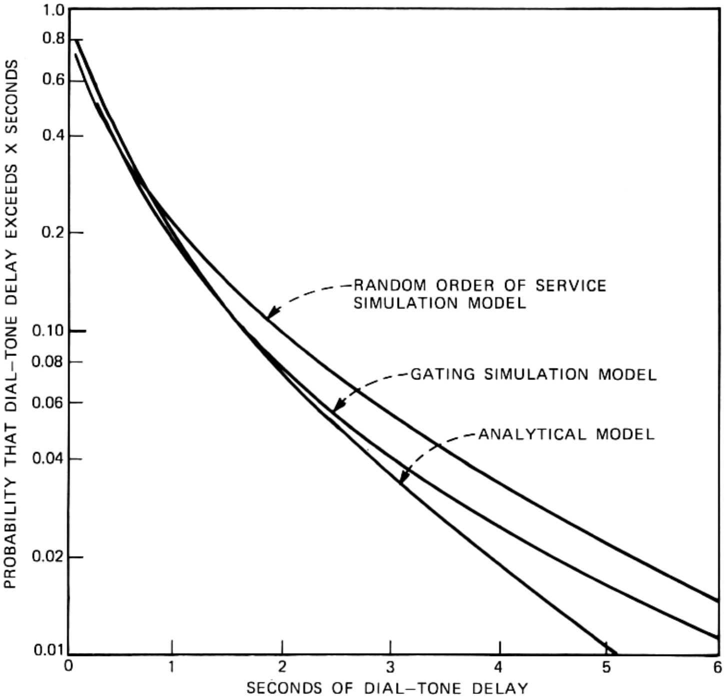

Fig. 2—Dial-tone delays are at 1600 CCS/LLF and 0.80 DTM OCC (excluding DTM2FTMS).

"feedback" (occurring when an all-ORs-busy condition is encountered), and k blocked calls residing permanently in the system. The DTMs cannot distinguish these k permanently blocked bad calls from the good calls in the system, so each of the k bad calls and each of the good calls in the system compete on an equal basis for DTMs. Upon completion of service by a DTM, a bad call rejoins the queue. With probability q, the good calls served at any given time period rejoin the queue and with probability p they leave it. We will show that each of these queuing systems has the same saturation load and we will describe how the equilibrium-queue-length distribution and the equilibrium-waiting-time distribution for each system is calculated.

Viewing the whole process again from the standpoint of a Markov chain with finitely many noncommunicating classes, the queue-length and waiting-time distributions referred to above may be regarded as being conditional on k, the number of blocked calls (permanently) present in the system. Now let k be a random variable with a truncated Poisson distribution of mean x, or, equivalently, regard the Markov chain as having any initial distribution $\pi(k, i)$, where i denotes the number of good calls in the system and where the marginal distribution $\pi(k, \cdot)$ of the number of bad calls in the system is truncated Poisson with mean x. Then the (unconditional) equilibrium-queue-length and equilibrium-waiting-time distributions for this system may be computed by taking the expectation (with respect to the truncated Poisson distribution of k) of the conditional queue-length and waiting-time distributions for each of the individual systems represented by the classes C_k.

Let X_n denote the number of good calls in the queue at time nT and let Y denote the number of bad calls (permanently) present in the queuing system. Let

$$P_k(i, j) = \Pr [X_{n+1} = j | X_n = i \text{ and } Y = k]$$

and let $P_k(i)$ be the equilibrium-queue-length distribution for the case of k bad calls. Let $A(n)$ be the probability that n good calls arrive in one service time interval (of length T). Then, by assumption,

$$\begin{aligned} A(n) &= \frac{e^{-\lambda}\lambda^n}{n!} \qquad n = 0, 1, 2, \cdots \\ &= 0 \qquad\qquad \text{otherwise,} \end{aligned}$$

where

$\lambda = \lambda_1 NT$

$\lambda_1 =$ LLF originating call rate (calls per second on one LLF),

$N =$ number of LLFs served by the two DTMs.

The transition functions $P_k(\cdot, \cdot)$ are given by

$$\begin{aligned} P_0(0, j) &= A(j) \\ P_0(1, j) &= pA(j) + qA(j-1) \\ P_0(i, j) &= pA(j-i+2) + qA(j-i) \qquad \text{for } i \geqq 2, \end{aligned} \tag{1}$$

and for $k \geqq 1$,

$$\begin{aligned} &P_k(0, j) = A(j) \\ &P_k(i, j) \\ &\quad = p[k(k-1)A(j-i)+2ikA(j-i+1)+i(i-1)A(j-i+2)]/ \\ &\qquad\qquad [(k+i)(k+i-1)] + qA(j-i) \qquad \text{for} \quad i \geqq 1. \end{aligned} \tag{2}$$

Note that $P_k(i, j) = 0$ for $i > j+2$ and that $P_k(j+2, j) > 0$.

If the equilibrium distributions $P_k(\cdot)$ exist, then they satisfy the equation

$$P_k(j) = \sum_{i=0}^{j+2} P_k(i) P_k(i, j). \tag{3}$$

Let f_k be the generating function for the kth queue length process; then

$$f_k(z) = \sum_{i=0}^{\infty} P_k(i) z^i.$$

The generating function f_0 is easily seen to be given by

$$f_0(z) = \frac{p(z-1)[(z+1)P_0(0) + zP_0(1)]}{z^2 e^{\lambda(1-z)} - (p + qz^2)}. \tag{4}$$

If $P_0(\cdot)$ exists, then $\lim_{z \uparrow 1} f_0(z) = 1$, and hence

$$P_0(0) + \tfrac{1}{2} P_0(1) = 1 - \frac{\lambda}{2p}. \tag{5}$$

Let $S_k^{(2)}(p)$ denote the saturation load of the kth process with two DTMs, as discussed above, and let $S_k^{(1)}(p)$ denote the saturation load of the analogous one-DTM system with k blocked calls. It can be shown that $S_k^{(1)}(p) = p$ for all $k = 0, 1, 2, \cdots$, i.e., a necessary and sufficient condition that a steady-state queue length distribution exists in the one-DTM system is that $\lambda < p$.* Using this result, we will show that $S_k^{(2)}(p) = 2p$ for all $k = 0, 1, \cdots$.

Since the Markov chain for the kth process above is irreducible, $P_0(0) > 0$ and $P_0(1) > 0$ whenever the stationary distribution $P_0(\cdot)$ exists. Hence, it follows from eq. (5) that $S_0^{(2)}(p) \leqq 2p$. Since $S_{k_2}^{(2)}(p)$

* This result can be proved using a theorem of Kushner (Ref. 4) and some properties of generating functions.

$\leqq S_{k_1}^{(2)}(p) \leqq S_0^{(2)}(p) \leqq 2p$ for $k_2 \geqq k_1 \geqq 0$, it suffices to show that $S_k^{(2)}(p) \geqq 2p$ for all $k \geqq 0$.

Consider a two-DTM system with $2k$ blocked calls, which differs from the $2k$th process defined above only in that (*i*) each DTM maintains a separate queue with half the originating traffic being assigned to one of the DTMs and the other half being assigned to the other DTM and (*ii*) each DTM serves k blocked calls. Let $\tilde{S}_{2k}(p)$ denote the saturation load of this modified system. Then clearly, $\tilde{S}_{2k}(p) \leqq S_{2k}^{(2)}(p)$ and, by symmetry, $\tilde{S}_{2k}(p) = 2S_k^{(1)}(p) = 2p$, because the modified system simply consists of two identical one-marker systems working separately, with each being assigned half the incoming traffic. Hence, $S_{2k}^{(2)}(p) \geqq 2p$ and so $S_k^{(2)}(p) = 2p$ for all $k \geqq 0$.

Thus, for $\lambda/2p < 1$, a stationary distribution exists for the kth process for each $k \geqq 0$. In the sequel, we will confine ourselves to the case where $\lambda/2p < 1$. In this case, f_0 exists and is analytic for $|z| < 1$ and continuous for $|z| = 1$, so the numerator of the expression on the right-hand side of eq. (4) must vanish for any $|z| \leqq 1$ for which the denominator vanishes. It is easy to see that the denominator has a single real root on the open interval $(-1, 0)$ and that $z = 1$ is a root. By applying Rouche's theorem to the functions $z^2 \exp[\lambda(1-z)]$ and $p + qz^2$ along the circle $|z| = 1 + \epsilon$, one can show that for $\lambda/2p < 1$ and for $\epsilon > 0$ sufficiently small (depending on λ and p), the expression $z^2 \exp[\lambda(1-z)] - (p + qz^2) = 0$ has exactly two roots in the open disc $|z| < 1 + \epsilon$. Thus, the root on $(-1, 0)$ and the root at 1 are the only roots within the closed unit disc and

$$(\xi + 1)P_0(0) + \xi P_0(1) = 0, \tag{6}$$

where ξ is the unique root of the expression

$$z^2 e^{\lambda(1-z)} - (p + qz^2) = 0, \qquad -1 < z < 0. \tag{7}$$

Solving eqs. (5) and (6) for $P_0(0)$ and $P_0(1)$ yields the unique solution

$$P_0(0) = \frac{-\xi(2p - \lambda)}{(1-\xi)p}$$

$$P_0(1) = \frac{(1+\xi)(2p-\lambda)}{(1-\xi)p}, \tag{8}$$

where ξ is defined by (7). The values $P_0(j)$ for $j \geqq 2$ can now be computed recursively from (3) using (8).

When $k > 0$, computation of $P_k(\cdot)$ by the method of generating functions involves numerical solution of some rather unwieldy differential equations. The following approach, which makes use of properties of recurrent Markov chains, is easier to implement computationally.

In Appendix A we show that, for $\lambda/2p < 1$, $P_k(\cdot)$ is given by $c(\boldsymbol{\alpha} + r\boldsymbol{\beta})$, where c and r are constants and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are the two (particular) eigenvectors, defined by (9) below, of the transition matrix $P_k(i, j)$.* We further show that the condition, $\alpha_j + r\beta_j \geqq 0$ for each j, determines r uniquely and provides an easy way to calculate r. Once r is obtained, the constant c is uniquely determined by the condition that $\sum_{j=0}^{\infty} P_k(j) = 1$. Since the quantities α_j and β_j can be computed recursively, the above observations lead to a simple algorithm for computing $P_k(\cdot)$. The results are given below, the proof is in Appendix A. Let

$$\alpha_0 = 1$$

$$\alpha_1 = 0$$

$$\alpha_{n+2} = \frac{\alpha_n - \sum_{j=0}^{n+1} \alpha_j P_k(j, n)}{P_k(n+2, n)} \qquad n \geqq 0 \tag{9a}$$

$$\beta_0 = 0$$

$$\beta_1 = 1$$

$$\beta_{n+2} = \frac{\beta_n - \sum_{j=0}^{n+1} \beta_j P_k(j, n)}{P_k(n+2, n)} \qquad n \geqq 0. \tag{9b}$$

It is shown in Appendix A that there is a unique constant r satisfying

$$m_n \leqq r \leqq M_n \qquad \text{all } n \geqq 3 \tag{10a}$$

$$\lim_{n\to\infty} m_n = r = \lim_{n\to\infty} M_n, \tag{10b}$$

where, for $n \geqq 3$, the increasing sequence m_n and the decreasing sequence M_n are defined by

$$\begin{aligned} m_n &= \max_{j \leqq n} \left[-\frac{\alpha_j}{\beta_j} : \alpha_j < 0, \beta_j > 0 \right] \\ M_n &= \min_{j \leqq n} \left[-\frac{\alpha_j}{\beta_j} : \alpha_j > 0, \beta_j < 0 \right]. \end{aligned} \tag{11}$$

Since the quantities m_n and M_n can be computed recursively, eqs. (10a), (10b), and (11) yield a well-defined algorithm for computing r.

Once r has been computed, $P_k(j)$ can be calculated from (3), setting $P_k(0) = c$, $P_k(1) = rc$ and determining c by the requirement that $\sum_{j=0}^{\infty} P_k(j) = 1$.

[*] Each of c, r, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ depends on k. For each integer $j \geqq 0$, α_j and β_j, denote the jth components of the eigenvectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

Since recursive calculation of $P_k(j)$ by (3) involves successive subtractions, it is essential to perform all computations in double precision.

In a No. 5 crossbar switching machine, the measured DTM occupancy during an hour is defined to be the total work time (in seconds) of all the DTMs divided by the product of the number of DTMs and the number of seconds in an hour. Thus, the measured occupancy is the average fraction of time that a given DTM is occupied. In the model the (steady-state) DTM occupancy is defined to be the limit, as n approaches infinity, of the expected fraction of time that a given DTM is occupied during the n work cycles of length T beginning at times $T, 2T, \cdots, nT$. This limit is equal to the probability that a given DTM is occupied on a work cycle beginning with the system in steady state. Hence,

$$\begin{aligned}\text{DTM occupancy} &= P\text{ (two or more calls are waiting for dial tone)}\\ &\qquad + \tfrac{1}{2}P\text{ (exactly one call is waiting for dial tone)}\\ &= 1 - P\text{ (no calls are waiting for dial tone)}\\ &\qquad - \tfrac{1}{2}P\text{ (exactly one call is waiting for dial tone).}\end{aligned}$$

As discussed earlier, the number of bad (blocked) calls waiting for dial tone at any given time has a truncated Poisson distribution in the model. Let

x = total erlangs of blocked calls on two DTMs

K = maximum possible number of blocked calls that can be in the system (waiting for dial tone) at any time*

$$c_{xK} = \sum_{k=0}^{K} \frac{x^k}{k\,!}.$$

Then the probability that k blocked calls are in the system is given by $(x^k/k\,!)c_{xK}^{-1}$. Let ρ_u denote the (steady state) DTM occupancy for the case of no blocked calls in the system and let ρ_b denote the (steady state) DTM occupancy including the effect of blocked calls. It follows from (5) that

$$\rho_u = \lambda/2p. \tag{12a}$$

Hence,

$$\begin{aligned}\rho_b &= 1 - c_{xK}^{-1}P_0(0) - \tfrac{1}{2}\left[c_{xK}^{-1}P_0(1) + xc_{xK}^{-1}P_1(0)\right]\\ &= 1 - c_{xK}^{-1}(1-\rho_u) - (x/2)c_{xK}^{-1}P_1(0).\end{aligned} \tag{12b}$$

2.3 Delay equations for the limiting model

In this section, we calculate the dial-tone delay probabilities for a call that arrives when the system (just prior to the arrival of the call)

* The truncation parameter K determines the number of conditional delay distributions to be calculated, so K should be taken to be no larger than needed to retain sufficient accuracy in the calculations. For values of x in the range of interest $e^{-x}c_{xK}$ differs from 1 by less than about 10^{-3} for $K = 5$; thus, for computational purposes, K may be taken to be 5.

is in steady state. Recall that the queuing model is in discrete time with arrivals at times nT and DTMs operating at times nT. Calls served on the DTM cycle beginning at nT either exit the system at time $(n+1)T$ or else return to the queue at time $(n+1)T$.

Thus, the steady-state probability that a call experiences a dial-tone delay of $(m+1)T$ seconds can be computed recursively in terms of the probability of a delay of mT seconds by straightforward conditional probability calculations involving matrix multiplications.

Let

Y = number of bad calls (permanently) present in the system
X_n = number of good calls in the queue at time nT
$\hat{X}$ = number of good calls in the queue immediately after an arrival when, just prior to the arrival, the system was in steady state.

Thus, $\hat{X}$ is the total queue length just after arrival of a call when the system is in steady state. Let

$$\hat{P}_k(j) = P\left\{X_n = j \;\middle|\; \begin{array}{l}\text{Queue is in steady state at time } (n-1)T \\ \quad\text{with } k \text{ bad calls permanently present in} \\ \quad\text{the system, and at least one good call} \\ \quad\text{arrives at time } nT\end{array}\right\}.$$

Then

$$P[\hat{X} = j \mid Y = k] = \hat{P}_k(j).$$

Let

$$\hat{P}_k(i, j) = P\left[X_{n+1} = j \;\middle|\; \begin{array}{l}X_n = i,\ Y = k, \text{ and at least one good} \\ \quad\text{call arrives at time } nT\end{array}\right]$$

and

$$\begin{aligned}\hat{A}(n) &= \frac{A(n)}{1 - A(0)} \qquad && n \geqq 1 \\ &= 0 && \text{otherwise.}\end{aligned}$$

Then $\hat{A}(n)$ is the conditional probability that n calls arrive at time nT given that at least one call arrives at time nT and $\hat{P}_k(i, j)$ is defined by eqs. (1) and (2) with $\hat{A}$ used in place of A. Also,

$$\hat{P}_k(j) = \sum_{i=0}^{j+2} P_k(i)\hat{P}_k(i, j).$$

Let $W_n(i, k)$ be the conditional probability that a call arriving in steady state has a dial-tone delay of nT seconds, given that the queue length of good calls upon arrival of the call is i and that the number of bad calls permanently present in the system is k. Thus, letting D

denote the dial-tone delay, we have

$$W_n(i, k) = P[D = nT \mid \hat{X} = i \text{ and } Y = k] \qquad i \geqq 1.$$

Then

$$W_1(i, k) = \min\left[1, \frac{2}{(k+i)}\right] p \qquad i \geqq 1,$$

and $W_{n+1}(\cdot, k)$ can be computed recursively from $W_n(\cdot, k)$ as described below.

Let $X_n = i \geqq 1$ and consider any one given call out of the i calls present at time nT. Let

$$\mathcal{P}_k(i, j) = P\left[\begin{array}{l|l} X_{n+1} = j \text{ and the given call does} & \\ \text{not leave the system on the DTM} & X_n = i \text{ and } Y = k \\ \text{cycle beginning at time } nT & \end{array}\right].$$

Then

$$W_{n+1}(i, k) = \sum_{j=0}^{\infty} \mathcal{P}_k(i, j) W_n(j, k),$$

and, for $i \geqq 1$, $\mathcal{P}_k(i, j)$ is given by*

$$\mathcal{P}_0(i, j) = \left[1 - \frac{\min(2, i)}{i}\right][pA(j - i + 2) + qA(j - i)] + \left[\frac{\min(2, i)}{i}\right] qA(j - i)$$

$$\mathcal{P}_1(i, j) = \left[\frac{i-1}{i+1}\right]\left[\frac{p}{i}[(i-2)A(j-i+2) + 2A(j-i+1)] + qA(j-i)\right] + \left[\frac{2}{1+i}\right] qA(j-i)$$

and, for $k \geqq 2$,

$$\mathcal{P}_k(i, j) = \left[1 - \frac{2}{k+i}\right]\left\{p \sum_{l=0}^{2} \frac{\binom{i-1}{l}\binom{k}{2-l}}{\binom{k+i-1}{2}} A(j - i + l) + qA(j - i)\right\} + \left(\frac{2}{k+i}\right) qA(j - i).$$

The second term on the right-hand side of the above equation is the probability of the event that: (*i*) the given call *is* selected for

[*] The derivation of $\mathcal{P}_k(i, j)$ for $k \geqq 2$ is given below. The derivations of $\mathcal{P}_k(i, j)$ for $k = 0$ and $k = 1$ are analogous and are omitted.

service by the DTMs, (*ii*) all ORs are busy, and (*iii*) $X_{n+1} = j$. (Recall that, in the model, either all ORs are busy, with probability q, or else two or more ORs are available, with probability $p = 1 - q$). The first term is the probability that: (*i*) the given call is *not* selected for service and (*ii*) $X_{n+1} = j$. In the first term, the expression in large brackets is the conditional probability that $X_{n+1} = j$ given that the call is not selected for service. This conditional probability is itself composed of the probabilities of two mutually exclusive events. The second term within the large brackets is the (conditional) probability of the event that all ORs are busy and $X_{n+1} = j$. The first term within the large brackets is the (conditional) probability that two or more ORs are free and $X_{n+1} = j$. The sum from $l = 0$ to $l = 2$ in this term pertains to the cases where 0, 1, or 2 good calls, respectively, are selected for service by the DTMs. Since this sum is part of a probability conditional upon a given good call not having been selected by the DTMs, the available population from which the DTMs may select calls consists of k bad calls ($k \geqq 2$) and $i - 1$ good calls ($i \geqq 1$). Since the selection is without replacement, the selection probabilities have the hypergeometric form shown above.

By the law of total probabilities,

$$P[D = nT | Y = k] = \sum_{i=1}^{\infty} W_n(i, k) P[\hat{X} = i | Y = k],$$

$$P[D > nT | Y = k] = 1 - \sum_{m=1}^{n} P[D = mT | Y = k],$$

and

$$P[D > nT] = \sum_{k=0}^{K} \frac{c_{xK}^{-1} x^k}{k!} P[D > nT | Y = k], \qquad (13)$$

where

$$c_{xK} = \sum_{k=0}^{K} x^k / k!.$$

III. CALCULATION OF OFFERED BLOCKED-CALL LOAD

In the limiting analytical model, the number of blocked calls in the system has a time-independent truncated Poisson distribution with mean x. This section describes a method for computing x, the mean offered erlangs of bad calls, in terms of the distribution of carried load among horizontal groups in the office. Using the limiting analytical model in conjunction with these methods for calculating the offered blocked-call load, we can calculate the No. 5 crossbar dial-tone delay distribution in terms of the calling rate and the distribution of carried load among horizontal groups in the office. Thus, we can predict the effect of frame load and balance on dial-tone delay in No. 5 crossbar offices.

We first construct a model for an individual horizontal group and express the expected bad-call load from one horizontal group as a function of the carried horizontal group load. To get the total expected bad-call offered load for the office, the expected contribution from an individual horizontal group is integrated (numerically) over the office horizontal group load distribution. The model, described below, for representing an individual horizontal group may be called a modified finite source Palm delay model.

We assume that the input to a horizontal group is from a finite number of sources with equal calling rates and that call-holding times are exponential. The number of sources N is taken to be 35 although in actuality most horizontal groups have 49 or 59 subscriber lines. The reason for the use of the lower number of sources is that an earlier study by W. S. Hayward, Jr.[5] showed that blocking on concentrators with unequal line occupancies can be approximated by blocking calculations based on equal calling rates and a lower number of sources. The use of 35 sources was suggested by J. G. Kappel.[6]

The calculations take into account the fact that an incoming call cannot occupy a subscriber line when all ten line links are occupied. Calls that find all line links busy will either defect or will eventually obtain a line link. While waiting for a line link to become available, a call is assumed to have an exponential waiting time until defection, with a mean of 30 s.*

In the case of ideal load balance, each horizontal group in the office is assumed to have a true carried load of $\bar{z}$ erlangs. In the case of less than ideal load balance, the distribution of true carried load among the individual horizontal groups is assumed to be normal with mean $\bar{z}$ and coefficient of variation c_g, where $\bar{z}$ is the office average carried horizontal group load and where c_g is the group-to-group coefficient of load variation for the office. The term c_g may be inferred from office load balance data either by using analysis of variance or, more commonly, by subtracting a standard value of the residual variance from the total measured variance of the office horizontal group load distribution.

* This is the same value that was used in step-by-step dial-tone-delay calculations. (See Ref. 7.) These calculations are based on the Palm delay model[8] using an assumed mean call holding time of 150 s and an assumed j factor of 5. In the notation of Ref. 9, this value of j corresponds to a mean-time-to-defection of 30 s. In Ref. 9, it is also stated that this value was found to be slightly conservative for most applications, based on review of panel office data reported in Ref. 9 and other (unpublished) step-by-step data.

To obtain the total offered blocked-call load for the office, we write the offered blocked-call load for an individual horizontal group $x(z)$ as a function of the individual horizontal group carried load z and integrate the function over the office distribution of carried load. We now describe how $x(z)$ is calculated.

Let N denote the number of subscriber lines per horizontal group. As discussed above, N has been taken to be 35 in all computations. Let λ_n denote the combined originating and terminating rate when n subscriber lines are occupied and let μ_n denote the subscriber line hang-up rate when n subscriber lines are occupied. Then

$$\begin{aligned} \lambda_n &= \lambda(N-n) && 0 \leqq n \leqq 9 \\ &= \frac{\lambda}{2}(N-n) && 10 \leqq n \leqq N-1 \end{aligned}$$

and

$$\begin{aligned} \mu_n &= n/H && 0 \leqq n \leqq 10 \\ &= 10/H + (n-10)/H_b && 11 \leqq n \leqq N, \end{aligned}$$

where H denotes the office average line-link holding time and H_b denotes the reciprocal of the defection rate for a call that is waiting for a line link to become available. As discussed above, H_b is taken to be 30 s based on results in Ref. 9.

The parameter λ (combined originating and terminating rate per unoccupied subscriber line) is an unknown whose value will be obtained from the horizontal group-carried load z. The factor of $\frac{1}{2}$ appearing in the definition of λ_n for $10 \leqq n \leqq N-1$ reflects the fact that, when all 10 line links are busy, an incoming call cannot cause a subscriber line to be occupied. The definition of the hang-up rate μ_n for $11 \leqq n \leqq N$ reflects the assumption that, when all 10 line links are occupied, the holding times of the subscriber lines for which no line links are available should be shorter than full call holding times.

For a horizontal group with carried load z, let

π_n = steady-state probability that n (out of N) subscriber lines are occupied.

Then, using standard methods for computing the steady-state distribution of a birth-and-death process,[10]

$$\pi_0 = c \text{ and } \pi_n = c\left(\frac{\lambda_0\cdots\lambda_{n-1}}{\mu_1\cdots\mu_n}\right) \quad \text{for } n \geqq 1,$$

where the constant c is determined by the requirement that

$$\sum_{n=0}^{N} \pi_n = 1.$$

Then the carried horizontal group load z is given by

$$z = \sum_{n=1}^{9} n \cdot \pi_n + 10 \sum_{n=10}^{N} \pi_n. \tag{14}$$

It is not hard to show that the carried horizontal group load z is a strictly increasing function of the subscriber line occupancy rate λ and that $z \leqq \min [10, \lambda NH]$. This makes it easy to determine numerically the unique value of λ corresponding to a given carried load z.

Once λ has been determined, the quantities λ_n, μ_n, and π_n are used to compute the offered blocked-call load contributed by the horizontal group. We take $x(z)$ to be the expected number of occupied subscriber lines for which no line links are available. This yields the value

$$x(z) = \sum_{n=11}^{N} (n - 10)\pi_n. \tag{15}$$

IV. PEAKEDNESS OF THE BLOCKED-CALL STREAM

In the model discussed in Section III, blocked calls arrive according to a Poisson process and defect after an exponential waiting time. Since blocked calls constitute an overflow stream and since it is well known[11] that overflow traffic usually has a peakedness* greater than 1 and hence is not Poisson, some discussion of the peakedness of the blocked-call stream is in order.

The blocked-call stream is the superposition of overflow traffic from all of the (typically 400 to 600) horizontal groups in an office. In the case of an office with ideal load balance (identical horizontal group loads), all horizontal groups would have equal expected contributions to the blocked-call stream and standard limit theorems would suggest that the blocked-call stream should be approximately Poisson.

Comparisons of calculated and observed dial-tone delays discussed in Section V (covering measured DTM occupancies up to about 0.84) indicate that, when the blocked-call stream is assumed to be Poisson, the calculated delays generally fall in the midrange of applicable data. However, there is a large variability in observed dial-tone delays measured in the same office under nearly identical levels of DTM occupancy, second-failures-to-match, and percent all-ORs-busy. The presence of this variability may be regarded as evidence that, in some busy hours, the blocked-call stream may be peaked in nature. Peakedness of the blocked-call stream is capable of accounting for substanti-

* The peakedness of a stream of calls is, by definition, the variance-to-mean ratio of the number of busy servers when the stream is offered to an infinite group of servers with independent identically distributed exponential holding times.

ally higher calculated dial-tone delay at a given level of DTM occupancy, second-failures-to-match, and percent all-ORs-busy than would be calculated under the assumption that the blocked-call stream is Poisson.

One would expect the blocked-call stream to be peaked whenever most of the blocked calls are contributed by a small number of highly overloaded horizontal groups. Data on the horizontal group load distributions during individual busy hours are not available for the test discussed in Section V and would be impractical to obtain on an ongoing basis in any office. In the absence of data from which one could deduce directly the blocked-call peakedness, a treatment has been made using some results which R. I. Wilkinson obtained in the course of formulating his "equivalent random" method of characterizing overflow traffic.[11]

Wilkinson[11] assumes that traffic arrives and departs according to a birth-and-death process in which the arrival rate is increased whenever the number of calls in the system exceeds a nominal number and is decreased whenever the number of calls in the system is less than this nominal number. The equilibrium distribution of the number of calls in the system resulting from these assumptions is shown to be negative binomial and, hence, this distribution is completely determined by its mean and variance (or by its mean and peakedness). Wilkinson then shows that negative binomial distributions rather closely approximate true overflow distributions under a number of different conditions.

The effects of peakedness of the blocked-call stream on dial-tone delay were explored using the limiting analytical model of Section II by replacing the (truncated) Poisson distribution of blocked calls by a (truncated) negative binomial distribution of blocked calls. (Note that this does not require recomputing the conditional delay distributions.) In this manner, dial-tone delay distributions were calculated and compared with the observed dial-tone-delay distributions discussed in Section V, assuming peakedness values of 2 and 4. The resulting delay curves approximated the higher delays observed for given occupancy parameters. These calculations indicate that much of the observed variability in dial-tone delays under nearly identical load can be explained by peakedness of the blocked-call load.

It is not difficult to modify the birth-and-death model discussed in Appendix B so as to accommodate peaked blocked-call input using Wilkinson's method of representing this input.[12]

Thus, when a negative binomial blocked-call distribution is used in the limiting model, the resulting distribution of the good-call dial-tone delay may be regarded as the limit of a sequence of good-call-delay distributions corresponding to models with blocked-call input having the peaked form suggested by Wilkinson.[11] The limit is taken as the bad-call-arrival rates approach zero and the mean bad-call-waiting times approach infinity with their products approaching positive constants. Incorporation of the negative binomial distribution in the limiting model is accomplished simply by replacing the terms of the truncated Poisson distribution in eq. (13) with the corresponding terms of the negative binomial distribution,[13] using any convenient truncation.

V. COMPARISON OF CALCULATED AND OBSERVED DIAL-TONE DELAY DISTRIBUTIONS

5.1 Summary

Theoretical dial-tone-delay distributions, calculated using the analytical model discussed in Section II, were compared with dial-tone-delay measurements made in a field test. The test was conducted in a No. 5 crossbar office with 60 LLFs, 4 DTMs, 68 dial-pulse (DP) ORs, and 68 multifrequency (MF) ORs. The data discussed in this section are from the time period February through April, 1974.

The main conclusions of this study are that (*i*) the calculated delays generally fall in the midrange of applicable data, (*ii*) there is a large variability in observed dial-tone delays measured under nearly identical levels of DTM occupancy, second-failures-to-match, and percent of all-ORs-busy, and (*iii*) the field data show a clear increase in the ratio of waste DTM usage to total DTM usage as frame load increases. The observed amount of increase in waste DTM usage agrees with theoretical predictions.

In this section, the manner in which the dial-tone-delay measurements were taken is discussed and the method used to obtain the calculated delays is outlined. Next, some of the sources of variability in No. 5 crossbar dial-tone delays are identified and an explanation is given as to why a large variability in observed delays should be expected. Two data plots are given indicating, respectively, the effects of frame load on waste DTM usage due to second-failures-to-match and the effects of frame load on incoming-first-failures-to-match. Finally, the conclusions of the study are discussed.

5.2 Methods of obtaining the calculated and observed delay distributions

Hourly dial-tone-delay measurements were made by placing approximately 900 test calls per hour, using a standard 3-s dial-tone-delay testing machine which had been modified to record the proportion of test calls with delays exceeding X seconds, for $X = 0.5$, 1.0, 1.5, 2.0, 2.5, and 3.0.

The *observed actual* DTM *occupancy* $\hat{\rho}_b$ for a single hour is expressed in terms of the measured parameters by the formula:

$$\hat{\rho}_b = \frac{(\text{Total DTM peg count})(0.015) + (\text{Measured seconds of DTM usage})}{(4)(3600)}. \tag{16}$$

The term in eq. (16) involving the DTM peg count is to adjust for the seconds of DTM usage which DTM usage measuring devices do not record.

The (adjusted) DTM holding time for each hour was computed by A. R. Thorne from the DTM peg count, the all-ORs-busy peg count, the adjusted DTM usage [the latter of which comprises the numerator of eq. (16)], and an additional "light-traffic adjustment" (used whenever the observed actual DTM occupancy is below 0.80). Most of the adjusted DTM holding times are between 0.28 s and 0.31 s. A DTM holding time of 0.30 s is assumed in the theoretical dial-tone-delay distributions discussed in this section.

To compare predicted dial-tone delays with measured dial-tone delays, it was necessary to infer the observed increment in DTM occupancy due to DTM second-failures-to-match (DTM2FTM). This increment, denoted by Δ, is taken to be

$$\Delta = \frac{(\text{Total DTM2FTM peg count})(HBC + 0.015)}{(4)(3600)}, \tag{17}$$

where

HBC = DTM holding time during a second-failure-to-match (seconds).

A value of 0.40 s is used for HBC, based on data obtained during an earlier dial-tone-delay field test.[6] The *observed good-call* DTM *occupancy* $\hat{\rho}_u$ is defined to be

$$\hat{\rho}_u = \hat{\rho}_b - \Delta. \tag{18}$$

Dial-tone-delay distributions were calculated, using the analytical model, for a range of values of actual DTM occupancy ρ_b and good-call DTM occupancy ρ_u, where the parameters ρ_b and ρ_u are defined in Section II. The specific manner in which the distributions were calculated is discussed below. The results of these calculations were tabulated into a set of dial-tone-delay distributions, indexed by the pairs (ρ_b, ρ_u). To compare the calculated and observed dial-tone delays, measured dial-tone-delay distributions from data-collection hours with similar values of $\hat{\rho}_b$ and $\hat{\rho}_u$ were plotted on a graph along with one theoretical dial-tone-delay distribution selected from the tabulation discussed above, such that $\rho_b \approx \hat{\rho}_b$ and $\rho_u \approx \hat{\rho}_u$ hold for the values $\hat{\rho}_b$ and $\hat{\rho}_u$ corresponding to the observed delays shown on the graph.

These graphs are shown in Figs. 3, 4, and 5. Table I lists data pertaining to each of the observed delay curves on each of the graphs. At the top of each graph are listed the actual DTM occupancy ρ_b and the good-call DTM occupancy ρ_u used for the theoretical dial-tone-delay curve (the solid line) on the graph. Also listed are the ranges of the $\hat{\rho}_b$ and $\hat{\rho}_u$ values corresponding to the observed dial-tone-delay curves on the graph. The plotting symbols on the graphs indicate measured dial-tone delays. The dotted lines are smoothing curves fitted to the measured delays by the computer plotting routine used to draw the graphs.

The predicted dial-tone-delay distributions were obtained in several steps. First, values of the good-call origination rate per LLF λ_1 were

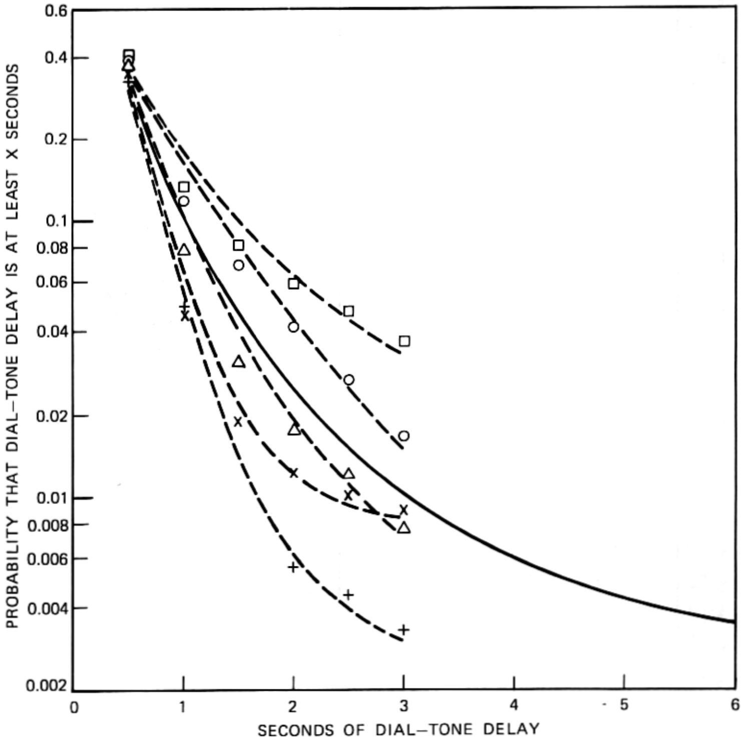

Fig. 3—Calculated and observed dial-tone delays. Actual occupancy in calculations = 0.75 (data: 0.740 to 0.758). Good-call occupancy in calculations = 0.70 (data: 0.684 to 0.697).

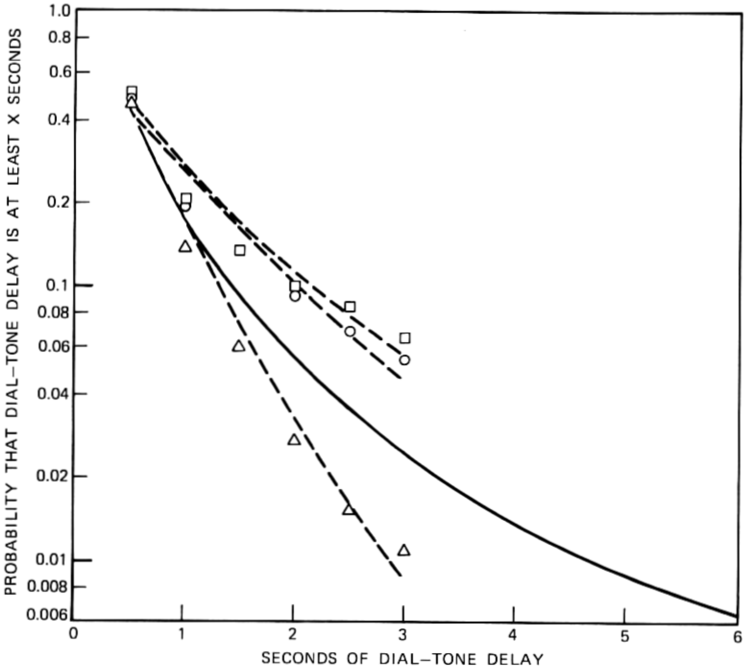

Fig. 4—Calculated and observed dial-tone delays. Actual occupancy in calculations = 0.82 (data: 0.795 to 0.837). Good-call occupancy in calculations = 0.76 (data: 0.744 to 0.782).

computed so as to produce good-call DTM occupancies of $\rho_u = 0.60$, 0.66, 0.68, 0.70, 0.72, 0.74, 0.76, 0.78, and 0.80. The values of λ_1 were obtained numerically from the formula $\rho_u = \lambda_1 NT/2p$ given in Section III. Note that p, the erlang C probability of all-ORs-busy, is a function of λ_1, whereas N and T are constants. For the office in which the test was conducted, $N = 30$ and, as discussed above, $T = 0.30$. In computing p, an average OR holding time of 10.25 s was assumed based on data from the test. In all cases, the calculated values of p were greater than 0.99, so the all-ORs-busy condition has a calculated probability of less than 0.01 under the conditions to which these distributions apply. (The observed fractions of all-ORs-busy were also below 0.01 during most of the hours of the test. Hence, ORs do not appear to have caused much of the dial-tone delay observed in the test.)

Next, the conditional delay distributions corresponding to these parameters were computed using the analytical model. Using eq. (12), values of x (total erlangs of blocked calls) corresponding to a range

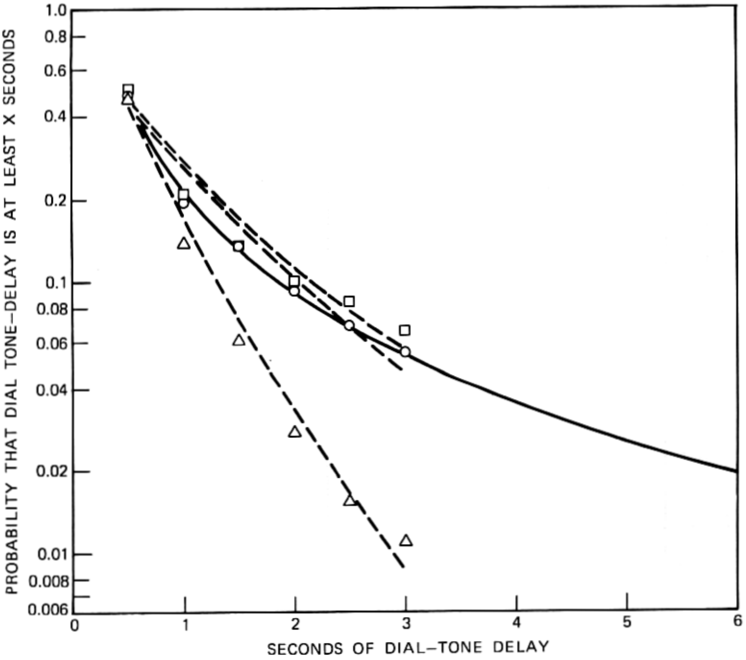

Fig. 5—Calculated and observed dial-tone delays. Actual occupancy in calculations = 0.82 (data: 0.795 to 0.837). Good-call occupancy in calculations = 0.76 (data: 0.744 to 0.782). (Note that a blocked-call peakedness of 4 is assumed in the calculations shown above. Figures 3 and 4 are based on an assumed blocked-call peakedness of 1.)

of values of ρ_b were computed for each fixed value of ρ_u. For each such pair of ρ_b and ρ_u, the good-call dial-tone-delay distributions were then computed by (13) using the value of x, obtained as discussed above, and the conditional delay distributions corresponding to ρ_u. The result of these calculations is a set of good-call dial-tone-delay distributions indexed by the pairs (ρ_b, ρ_u).

These good-call dial-tone-delay distributions include the *indirect* effect of bad calls in that they reflect the increased DTM congestion produced by the bad calls. As discussed earlier, the *direct* effect of bad calls is expected to be small and therefore may be accounted for in a somewhat approximate manner. To reduce the number of variables that need to be considered, a single blocking probability is used in lieu of integrating the blocking probability formula over an assumed distribution of expected horizontal group loads. For the purpose of computing an average blocking probability PB, the average waiting-

Table I — Dial-tone-delay data used for comparison with calculated delays

Fig. No.	Date	Time	% Actual DTM Occupancy	% Good-Call Occupancy	No. DTS Test Calls	% AORB	CCS/LLF	0.5 s*	1.0 s	1.5 s	2.0 s	2.5 s	3.0 s
3	3/4/74	1500–1600	74.0	69.7	900	0.33	N.A.	370	120	73	53	42	33
	3/25/74	1500–1600	74.5	68.4	900	0.30	1344	354	106	62	37	24	15
	4/9/74	1530–1630	75.5	69.4	900	0.00	1315	298	44	9	5	4	3
	2/6/74	1500–1600	75.8	69.7	900	0.01	1438	314	41	17	11	9	8
	2/22/74	1600–1700	74.7	69.4	900	0.06	1351	341	71	28	16	11	7
4 and 5	2/6/74	1600–1700	80.9	74.4	900	0.61	1480	461	188	122	90	76	59
	2/15/74	1600–1700	80.4	76.1	900	0.92	1390	430	175	121	83	62	49
	4/8/74	1530–1630	83.7	78.2	900	0.25	1384	426	126	53	25	14	10

* For each hour listed, the number in this column is the observed number of dial-tone speed (DTS) test calls whose dial-tone delay exceeds 0.5 s. The columns labeled "1.0 s" through "3.0 s" are defined analogously.

time-until-defection of a bad call (BCWT) is assumed to be 10 s.[*] A consequence of this assumption is that the total bad-call origination rate corresponding to x total erlangs of bad calls on two DTMs is $x/10$ bad calls per second. The total origination rate of good calls on all 30 LLFs is $30\lambda_1$. The average blocking probability PB corresponding to the occupancy pair (ρ_b, ρ_u) is then defined to be the ratio of the bad-call origination rate to the total origination rate of good and bad calls. Thus,

$$\text{PB} = \frac{x/10}{x/(10) + 30\lambda_1}, \tag{19}$$

where x and λ_1 are the values corresponding to the occupancy pair (ρ_b, ρ_u).

The conditional delay distribution of bad calls should be nearly exponential with mean delay $\text{HT}/(10 - a)$, where HT is the average call holding time and a is the carried horizontal group load in erlangs.[†] For each occupancy pair (ρ_b, ρ_u), the theoretical dial-tone-delay distribution for all calls is then given by

$$P(D > t) = (1 - \text{PB})P_G(D > t) + \text{PB}\left\{\exp\left[-\frac{(10 - a)\,t}{\text{HT}}\right]\right\}, \tag{20}$$

where $P_G(D > t)$ is the good-call dial-tone-delay distribution corresponding to the occupancy pair (ρ_b, ρ_u).

In the theoretical dial-tone-delay curves shown in Figs. 3, 4, and 5, $a = 0$ is used in eq. (20). The effect on $P(D > 3 \text{ s})$ of setting $a = 0$, rather than using a more nearly correct value for each graph, is less

[*] The value of BCWT = 10 s used in calculating PB is consistent with the mean-time-to-defection $H_b = 30$ s used in the horizontal group blocking calculations in Section III. The difference between the numerical values arises because these two parameters are defined differently. For the purpose of calculating PB, it is assumed that a bad call arrives, remains waiting for dial tone an exponential length of time with mean BCWT, and then defects. In the horizontal group blocking model discussed in Section III, a call which finds all 10 line links busy may either defect or may eventually obtain an idle line link. The call contributes to the bad-call load during the time that it remains waiting for one of the line links to become available. Thus, the quantity in the horizontal group blocking model corresponding most nearly to the parameter BCWT is the mean time until a call that finds all 10 line links busy either obtains a line link or defects. Based on delay calculations for the horizontal group blocking model discussed in Section III, and assuming a mean-time-to-defection of 30 s for calls that do not obtain a line link and a mean line-link holding time of 150 s, the mean time until a call that finds all 10 line links busy either obtains a line link or else defects is calculated to be between about 10 and 12 s for horizontal group carried loads in the range of interest. Thus, it is reasonable to take BCWT = 10 s.

[†] This expression for the conditional delay distribution of bad calls should overestimate their delays somewhat because the expression does not account for finite-source effects. Since the fraction of calls experiencing these delays (i.e., the fraction of bad calls) is typically less than 0.01, the effect on 3.0-s dial-tone-delay probabilities of neglecting finite-source effects is typically less than 5×10^{-4}. For this reason, these effects are neglected in eq. (20).

than 8×10^{-4} in all cases. This error is in the opposite direction from the error, of comparable magnitude, resulting from not including finite-source effects in the bad-call delay distribution.

Since the analytical model is a discrete time model with a time step of T seconds, $P(D > kT) = P[D \geqq (k+1)T]$ for $k = 1, 2, \cdots$. In Figs. 3, 4, and 5, $T = 0.30$. Because of the way that dial-tone-delay measurements are taken, a call which is recorded as having a delay greater than t seconds may actually receive dial tone within a few milliseconds after time t. Thus, in comparing the theoretical dial-tone delays with the observed dial-tone delays, the observed fraction of dial-tone delays *greater than* t seconds is taken as representing the observed fraction of dial-tone delays *greater than or equal to* t seconds. The theoretical delay curve plotted is the curve $P(D \geqq kT)$ for $k = 1, 2, \cdots$, interpolated so as to produce a smooth curve.

5.3 Sources of variation in observed No. 5 crossbar dial-tone delays

Dial-tone delay in No. 5 crossbar is influenced by a number of factors capable of producing a large variation in delays measured in different hours within the same office under very similar conditions of DTM occupancy, percent all-ORs-busy, and second-failures-to-match. As discussed in Section IV, much of the variability in dial-tone delay measured under very similar load conditions can be explained by differences in peakedness of the blocked call stream. Whenever most of the blocked-call load comes from a small number of extremely overloaded horizontal groups, the blocked-call stream should have a peakedness greater than one. When a large number of moderately overloaded groups contribute to the blocked-call load, the blocked-call stream should be approximately Poisson (peakedness equal to one). Thus, differences in the individual busy-hour-load balance would be expected to produce different amounts of blocked-call peakedness, which in turn can account for appreciable differences in dial-tone delay measured under nearly identical average load conditions. For example, Fig. 5 shows dial-tone delays calculated assuming a blocked-call peakedness of 4 for the same conditions shown in Fig. 4, which is based on a blocked-call peakedness of 1.* The calculated delay curve in Fig. 5 fits the top two observed delay distributions rather closely.

Some additional identifiable sources of dial-tone-delay variation under similar load conditions are: (*i*) within-hour trends in traffic, (*ii*) nonstandard (and possibly erratic) gating caused by improper functioning of the master traffic controller circuitry, (*iii*) DTM prefer-

* The method by which blocked-call peakedness is treated in the model is discussed in Section IV.

ence for calls from a small subset of lines on each horizontal group, (*iv*) variation in DTM first-failures-to-match, and (*v*) competition between DTMs and completing markers for line-link connectors.

The first of these sources should produce effects similar to those of blocked-call peakedness. The second source may cause nonuniform congestion. The third source is predicted to result in a slight outward shift in the delay curve. The fourth and fifth sources should be reflected in increased measured DTM holding time and in increased DTM occupancy. Although approximate allowances can be made for the average congestion increase produced by some of these phenomena, no quantitative estimate is available for their total contribution to hourly *variation* in dial-tone delay.

In addition to identifiable sources of dial-tone-delay variation, simulation studies indicate that there can be an appreciable residual variation in simulated hourly No. 5 crossbar dial-tone delays obtained in different runs with identical input parameters (and, hence, with identical expected load conditions).*

Figure 6 shows four dial-tone-delay distributions obtained using the gating simulation model. In this model, the blocked-call stream is Poisson. These distributions were produced by simulating four individual hours, using identical input parameters. The delay distributions shown are for the calls that did not encounter horizontal group blocking and are based on the total number of such calls processed during the hour. The set of four 3-s dial-tone delays has a coefficient of variation of 0.28 and a mean of 0.069. (The coefficient of variation is the ratio of the standard deviation to the mean.) Plotted on the graph along with the delay curves are error bars indicating the 2-sigma limits of 0.034 and 0.107 associated with the above mean and coefficient of variation.

Actual dial-tone-delay measurements are based on test calls. During a given busy hour in a typical No. 5 crossbar office, approximately 900 test calls are made on a fixed set of 60 (out of 600) horizontal groups. The use of test calls introduces sampling error, which is not represented in the distributions shown in Fig. 6 and which would have an associated coefficient of variation of about 0.12 for the parameters applicable to Fig. 6.

5.4 Data on frame load effects

Figure 7 is a data plot of line-link frame load versus waste DTM usage due to second-failures-to-match based on data from the test. The quantity "DTM usage fraction due to 2FTMs" shown on the ordinate is

* This result is one of the main conclusions of an earlier No. 5 crossbar dial-tone-delay simulation study conducted by S. Halfin.[3]

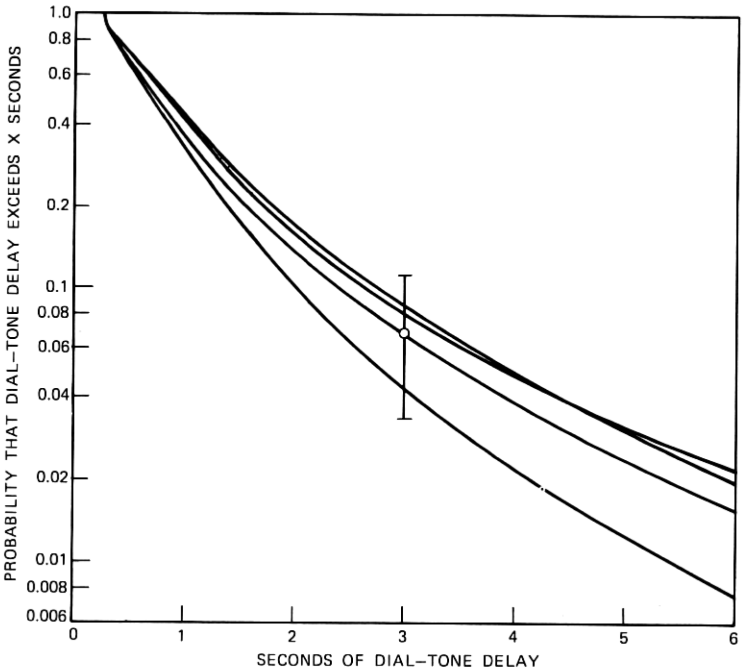

Fig. 6—Simulated No. 5 crossbar dial-tone delays. Distributions are based on simulation of four individual hours with identical inputs.

defined by

$$\text{DTM usage fraction due to 2FTMs} = \frac{\text{Total DTM usage due to 2FTMs}}{\text{Total DTM usage}}$$

$$= \frac{\Delta}{\hat{\rho}_b}.$$

Figure 8 is a data plot of line-link frame load versus incoming-first-failure-to-match (IFFM) based on data from the test. In each figure, each of the data points represents data from 1 h. The average actual DTM occupancy (averaged over all the data-collection hours with measured line-link frame loads of 1100 CCS/LLF or more) is 0.54; the DTM occupancy range is from less than 0.40 to 0.84.

Existing theory indicates that, for a given office configuration (including a given junctor pattern), IFFM is directly and primarily dependent on frame load. This conclusion is borne out by Fig. 8, which demonstrates a well-defined trend (with a moderate amount of data scatter) of increasing IFFM with increasing frame load.

The extent to which frame load affects DTM usage and dial-tone delay in any given hour depends, in an indirect way, on several differ-

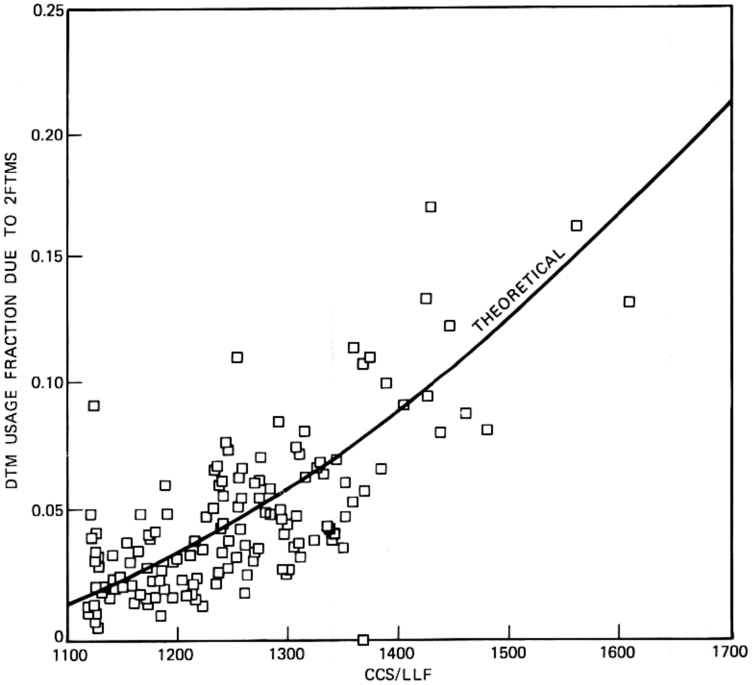

Fig. 7—Effect of frame load on increased DTM usage due to 2FTMs. The assumed DTM holding time during a 2FTM = 0.40 s.

ent variables, including DTM occupancy and the distribution of carried load among horizontal groups. Because of this dependence, the nature of which has been discussed more fully in previous sections, the data scatter in the plot of waste DTM usage fraction versus frame load (Fig. 7) is much larger than in the plot of IFFM versus frame load (Fig. 8).

The curve labeled "THEORETICAL" in Fig. 7 was calculated using the limiting analytical model [eq. (12)] and the horizontal group blocking model discussed in Section III. In these calculations, the blocked-call load was assumed to be Poisson (peakedness equal to 1).

To obtain the theoretical curve, it was first necessary to determine what calling rates should be assumed in the calculations. The calling rates were inferred from the data upon which Fig. 7 is based by first using linear regression (least squares) to express the observed good-call DTM occupancy, $\hat{\rho}_u$, as an empirical function of the observed frame load, CCS/LLF. For each value of the frame load, the calling rate was taken to be the particular calling rate corresponding to the least squares value of $\hat{\rho}_u$, assuming an average DTM holding time of 0.30 s and a

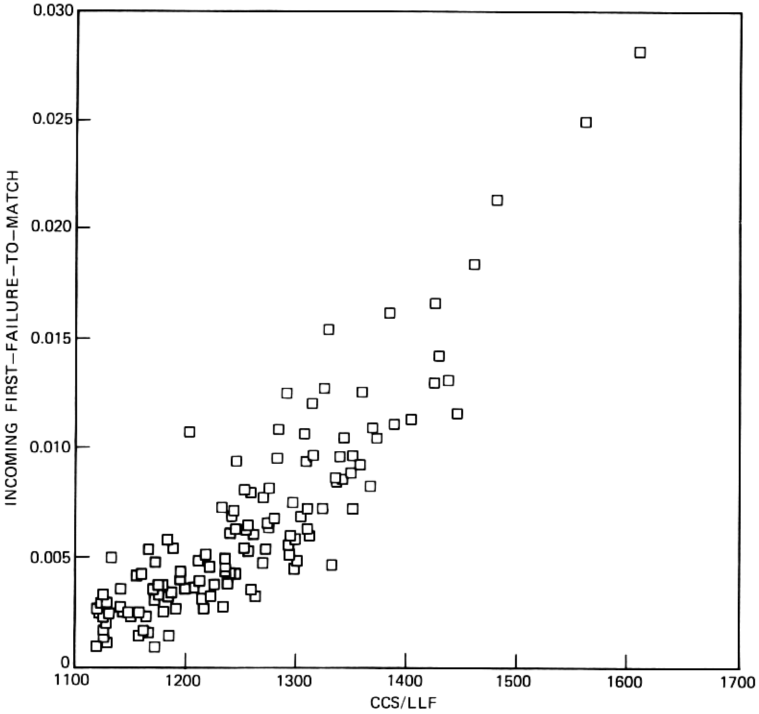

Fig. 8—Effect of frame load on incoming first-failure-to-match.

zero probability of all-ORs-busy. The justification for the assumptions regarding the DTM holding time and the probability of all-ORs-busy is given earlier in this section.

Next, the horizontal group blocking model discussed in Section III was used to compute the expected blocked-call offered load corresponding to each value of frame load and (empirically associated) calling rate. The office distribution of carried horizontal group load was assumed to be normal with group-to-group coefficient of variation inferred from the office horizontal group load distribution using the method discussed earlier. The calculated values of offered blocked-call load were then used in eq. (12) to compute the theoretically predicted fraction of waste DTM use, $(\rho_b - \rho_u)/\rho_b$, corresponding to these frame loads and calling rates.

Since both the theoretical and the observed values of waste DTM use depend not only on frame load but also on the calling rate, neither the data plotted in Fig. 7 nor the theoretical curve shown on the figure should be regarded as being applicable to calling rates or load balance conditions other than those upon which this figure is based.

The purpose of Fig. 7 is to illustrate—for the conditions of calling rate, frame load, and load balance represented in this study—an empirical relationship between frame load and the ratio of waste DTM usage to total DTM usage with increasing frame load and to show that this empirical relationship is consistent with predictions of the theoretical models.

5.5 Conclusions

The conclusions which the author has drawn from this comparison of theoretical and observed dial-tone-delay distributions are as follows:

(*i*) Where two or more observed delay distributions appear on the same graph, the theoretical delay distributions usually fall approximately midway between the maximum and minimum observed delay distributions. This indicates that, in the (actual) DTM occupancy range spanned by these data (DTM occupancies up to about 0.84), the analytical model shows good agreement with the data.

(*ii*) Much of the large variability in observed 3-s dial-tone delays measured under nearly equal conditions of DTM occupancy, second-failures-to-match, and percent all-ORs-busy can be explained by assuming the blocked-call stream to have different peakedness values (ranging from 1 to about 4) in different hours. High blocked-call peakedness, illustrated by Fig. 5 in which the peakedness is taken to be four, would be expected whenever most of the blocked-call load comes from a small number of extremely overloaded horizontal groups. Low blocked-call peakedness, illustrated by Fig. 4 in which the peakedness is taken to be 1, would be expected whenever a large number of moderately overloaded horizontal groups contribute more or less equally to the blocked-call load.

(*iii*) Simulation results indicate that there should be a large residual variability in 3-s dial-tone delays measured in different (simulated) hours under identical expected load conditions. This variability, illustrated in Fig. 6, is in addition to the variability due to blocked-call peakedness discussed above. A third nonnegligible source of variability in observed dial-tone delays is the sampling variability associated with the use of test calls to measure delays.

(*iv*) For the conditions of calling rate, frame load, and load balance represented in this study, the observed increase in the ratio of waste DTM usage to total DTM usage as frame load increases is consistent with theoretical predictions. This is illustrated in Fig. 7 and discussed in Section 5.4.

VI. ACKNOWLEDGMENT

The author would especially like to thank J. G. Kappel for guiding this work and contributing numerous suggestions. In addition, the author appreciates useful comments, results, and suggestions contributed by A. L. Dudick, S. Halfin, W. S. Hayward, Jr., N. A. Marlow, A. R. Thorne, and F. R. Wallace of Bell Laboratories; H. S. McKane and H. A. Reading of AT&T; and J. C. Warner of Michigan Bell Telephone Company.

APPENDIX A

Computation of the Two-Marker Stationary Distribution for $k \geqq 1$

The purpose of this appendix is to prove that the limit r defined by (10) exists and that the quantities $v_j = \alpha_j + r\beta_j$, satisfy (21), where α_j and β_j are defined by (9).

$$\begin{aligned} v_j &\geqq 0, && \text{all } j \geqq 0 \\ v_0 &= 1; \\ v_j &= \sum_{i=0}^{j+2} v_i P_k(i, j), && \text{all } j \geqq 0 \\ &\sum_{j=0}^{\infty} v_j < \infty. \end{aligned} \tag{21}$$

We first show that there exists exactly one positive number r such that $v_j = \alpha_j + r\beta_j$ satisfies (21). Then we show that the limit in (10) exists and is equal to this number.

When $\lambda < 2p$, we know that the stationary distribution P_k exists and $P_k(j) > 0$ for each j. Let $\pi_j = P_k(j)/P_k(0)$. Then for $r = P_k(1)/P_k(0)$, we have

$$\begin{aligned} \pi_0 &= 1, \\ \pi_1 &= r, \\ \pi_j &= \sum_{i=0}^{j+2} \pi_i P_k(i, j), \qquad \text{all } j \geqq 0. \end{aligned}$$

Now let $v_j = \alpha_j + r\beta_j$, with r as defined above. Then, it follows from (9) that

$$\begin{aligned} v_0 &= 1, \\ v_1 &= r, \\ v_j &= \sum_{i=0}^{j+2} v_i P_k(i, j), \qquad \text{all } j \geqq 0. \end{aligned} \tag{22}$$

Since $v_0 = \pi_0$ and $v_1 = \pi_1$ and since the vectors $\mathbf{v}$ and $\boldsymbol{\pi}$ both satisfy the same recurrence relation, which has the property that the jth term for each $j \geqq 2$ is uniquely determined by the 0th and 1st terms, we see that $v_j = \pi_j > 0$ for all $j \geqq 0$. Since the Markov chain is positive recurrent for $\lambda/2p < 1$, it follows from (22) and Karlin[14] that $\sum_{j=0}^{\infty} v_j < \infty$. Hence, $\mathbf{v}$ satisfies (21). Thus, there exists at least one value of r for which $v_j = \alpha_j + r\beta_j$ satisfies (21). If r' is any number such that $v'_j = \alpha_j + r'\beta_j$ satisfies (21), then by uniqueness of the stationary distribution, $r' = v'_1/v'_0 = P_k(1)/P_k(0)$, so $r' = r$. Thus, exactly one such r exists.

Since $\alpha_j + r\beta_j > 0$ for each j, it follows that (10a) holds. Hence, the increasing sequence m_n is bounded above by r and the decreasing sequence M_n is bounded below by r. So both $m \equiv \lim_n m_n$ and $M \equiv \lim_n M_n$ exist and satisfy

$$m \leqq r \leqq M, \tag{23}$$

$$\begin{aligned} m_n &\leqq m, \\ M &\leqq M_n. \end{aligned} \tag{24}$$

If either of the inequalities in (23) were strict, then $m < M$ and, in view of (24), each x in the interval $m < x < M$ would satisfy

$$m_n \leqq x \leqq M_n$$

for each n, from which it follows that $v_j = \alpha_j + x\beta_j$ satisfies (21). This contradicts the uniqueness of r and proves that $m = M = r$.

It is easy to extend the above result so as to give a method for computing the stationary distribution of any irreducible positive recurrent Markov chain on the nonnegative integers such that a fixed positive integer n_0 exists for which $P(j + n_0, j) > 0$ and $P(j + n, j) = 0$ for all j and all $n > n_0$.

APPENDIX B

Convergence in Distribution to the Limiting Model

In the limiting model, the number of bad (i.e., blocked) calls in the system is a time-independent, truncated-Poisson-distributed random variable. We will show that the steady-state distributions of good- and bad-call queue lengths and the steady-state good-call dial-tone-delay distribution in the limiting model are the limits of the corresponding distributions for a sequence of models in which the bad call queue lengths form birth-and-death processes.

In the nth model, the bad-call arrival rate, denoted by $\lambda^{(n)}$, approaches zero and the bad-call mean waiting time until defection, denoted by $H^{(n)}$, approaches infinity in such a way that their product (total expected erlangs of bad calls) approaches a finite positive constant x.

The physical motivation for considering the limiting model is that the actual situation is one in which blocked calls appear infrequently (relative to total call arrival rates) and tend to remain in the system for a long time (relative to DTM holding times). The mathematical motivation is that the limiting model is much easier to analyze than a model in which the bad-call arrival and departure processes are represented explicitly.

Preliminary computations using a single marker version of the analytical model, which represented the bad-call arrival and departure processes explicitly, showed that different bad-call arrival rates and waiting times had little effect on dial-tone-delay distributions as long as the product of the bad-call arrival rate and bad-call waiting time remained constant. In addition, these distributions were all quite close to those obtained from the limiting form of the model. These results make sense intuitively because bad calls simply cycle through the system, absorbing some DTM uses while present, and eventually defect; thus, what should matter is mainly the distribution of the number of bad calls in the system at any time. In the version of the analytical model discussed in this appendix, the blocked call queue length distribution is shown to be a truncated Poisson with mean equal to the total erlangs of bad calls.

Let K be the maximum possible number of bad calls that can be in the system at any time. In all computations using realistic No. 5 crossbar busy-hour input parameters, the expected erlangs of bad calls have been low enough that the Poisson probability of more than five blocked calls being present in a two-marker system has been less than 10^{-3}. Hence, K may be taken to be 5. (Note that in an actual No. 5 crossbar office, the maximum number of bad calls that can be in the system at any time is trivially bounded above by the number of subscriber line terminations in the office.)

Let $\hat{S}_n(p)$ denote the saturation load of the nth model. Since no more than K blocked calls can be in the queue at any time, in the nth model, we have $2p = S_0^{(2)}(p) \geqq \hat{S}_n(p) \geqq S_K^{(2)}(p) = 2p$ for all n; hence, $\hat{S}_n(p) = 2p$. Thus, the saturation load of the nth model is the same as the saturation load of the limiting model. For the nth model, let

$$\begin{aligned} X_m^{(n)} &= \text{number of good calls in the queue at time } mT, \\ Y^{(n)}(mT) &= \text{number of bad calls in the queue at time } mT, \\ p_n(k, l) &= P\{Y^{(n)}[(m+1)T] = l \,|\, Y^{(n)}(mT) = k\}. \end{aligned}$$

In the nth model, $Y^{(n)}(t)$, the bad-call population size at time t, is a finite-state space birth-and-death process with birth-and-death rates

$$\begin{aligned} \lambda_m^{(n)} &= \lambda^{(n)} && 0 \leqq m \leqq K-1 \\ &= 0 && \text{otherwise} \\ \mu_m^{(n)} &= \frac{m}{H^{(n)}} && 0 \leqq m \leqq K \\ &= 0 && \text{otherwise.} \end{aligned}$$

Since the queuing model is in discrete time, only the values of $Y^{(n)}(t)$ at $t = mT$ are of interest. The transition matrix $p_n(k, l)$ is given by

$$\mathbf{p} = e^{T\mathbf{A}_n},$$

where $\mathbf{A}_n$ is the infinitesimal generator matrix of the birth-and-death process $Y^{(n)}$. (See Ref. 15.) Since $\lim_{n\to\infty} \lambda^{(n)} = 0$ and $\lim_{n\to\infty} H^{(n)} = \infty$, it follows that $\lim_{n\to\infty} \mathbf{A}_n = \mathbf{0}$. Hence,

$$\lim_{n\to\infty} p_n(k, l) = \delta_{kl} \tag{25}$$

for each $k, l = 0, 1, \cdots, K$, where $\delta_{kl} = 0$ for $k \neq l$ and $\delta_{kk} = 1$.

We also have

$$P\{X_{m+1}^{(n)} = j, Y^{(n)}[(m+1)T] = l \mid X_m^{(n)} = i, Y_{(mT)}^{(n)} = k\} = p_n(k, l)P_k(i, j),$$

where $P_k(i, j)$ is given by (1) and (2). Since the nth model has saturation load $2p$, we know that a stationary queue length distribution $\pi_n(j, l)$ exists for $\lambda/2p < 1$ and satisfies the equation

$$\pi_n(j, l) = \sum_{k=0}^{K} \sum_{i=0}^{j+2} \pi_n(i, k)p_n(k, l)P_k(i, j). \tag{26}$$

Let $\pi_n(\cdot, l)$ be the marginal equilibrium distribution of the number of bad calls in the system. Then

$$\pi_n(\cdot, l) = \sum_{j=0}^{\infty} \pi_n(j, l) \tag{27}$$

and by (26)

$$\begin{aligned} \pi_n(\cdot, l) &= \sum_{k=0}^{K} \sum_{i=0}^{\infty} \pi_n(i, k)p_n(k, l) \sum_{j=0}^{\infty} P_k(i, j) \\ &= \sum_{k=0}^{K} \pi_n(\cdot, k)p_n(k, l). \end{aligned} \tag{28}$$

Hence, $\pi_n(\cdot, k)$ is the stationary distribution of a Markov chain whose transition matrix is $p_n(k, l)$. Using standard theorems on birth-and-death processes (Ref. 10) it follows that, for all values of $l = 0, 1, \cdots, K$,

$$\begin{aligned}\pi_n(\cdot, k) &= \lim_{t\to\infty} P[Y^{(n)}(t) = k \mid Y^{(n)}(0) = l] \\ &= \frac{c_{x_nK}^{-1} x_n^k}{k!} \qquad 0 \leqq k \leqq K \\ &= 0 \qquad \text{otherwise,}\end{aligned} \tag{29}$$

where

$$x_n = \lambda^{(n)} H^{(n)}$$

and

$$c_{x_nK} = \sum_{k=0}^{K} \frac{x_n^k}{k!}.$$

We will now show that

$$\lim_{n\to\infty} \pi_n(j, k) = \frac{c_{xK}^{-1} x^k}{k!} P_k(j) \tag{30}$$

for all $k = 0, 1, \cdots, K$ and all $j = 0, 1, \cdots$, where P_k is the stationary distribution of good-call queue length in the model with k bad calls permanently present in the system. The right-hand side of (30) is the queue-length distribution for good and bad calls in the limiting model.

Let $X^{(n)}$ denote the number of good calls in the queue for the nth model in equilibrium. Since the number of bad calls present in the nth model is always less than or equal to K for all n, it is clear on intuitive grounds and can be proved rigorously using stochastic ordering that

$$P[X^{(n)} \geqq j] \leqq \sum_{i=j}^{\infty} P_K(i) \tag{31}$$

for each j.

To prove (30), it suffices to show that if $\pi_{n'}$ is any subsequence of π_n for which $\lim_{n'} \pi_{n'}(j, k) \equiv \pi(j, k)$ exists for each j and k, then

$$(i) \quad \pi(j, k) = \sum_{i=0}^{j+2} \pi(i, k) P_k(i, j)$$

and

$$(ii) \quad \sum_{j=0}^{\infty} \pi(j, k) = \frac{c_{xK}^{-1} x^k}{k!}.$$

To see that (i) and (ii) are sufficient, note that, by uniqueness of the stationary distribution, P_k, (i) and (ii) imply

$$\pi(j, k) = \frac{c_{xK}^{-1} x^k}{k!} P_k(j). \tag{32}$$

Since $0 \leqq \pi_n(j, k) \leqq 1$ for all j and k, every subsequence $\pi_{n'}$ contains a further subsequence $\pi_{n''}$ for which the limit $\lim_{n''} \pi_{n''}(j, k)$ exists for all j and k. In view of the above discussion, all of these subsequences have the same limit, namely the right-hand side of (32). Equation (30) follows.

We now prove (*i*) and (*ii*). Condition (*i*) follows immediately from (25) and (26). To see that (*ii*) holds, we proceed in two steps. First, note that by (27), (29), and Fatou's lemma,

$$\sum_{j=0}^{\infty} \pi(j, k) \leqq \frac{c_{xK}^{-1} x^k}{k!}$$

for each k. If any of the above inequalities were strict, then

$$\sum_{k=0}^{K} \sum_{j=0}^{\infty} \pi(j, k) < 1.$$

Hence, to prove (*ii*) it suffices to show

$$\sum_{i=0}^{\infty} \sum_{k=0}^{K} \pi(i, k) \geqq 1,$$

i.e., we must show that, in the limit, no probability mass escapes to infinity. Let $\epsilon > 0$ and choose j_0 such that

$$\sum_{j=j_0}^{\infty} P_K(j) < \epsilon.$$

Then, using eq. (31),

$$\begin{aligned}\sum_{i=0}^{\infty} \sum_{k=0}^{K} \pi(i, k) &\geqq \sum_{i=0}^{j_0} \sum_{k=0}^{K} \pi(i, k)\\ &= \lim_{n''} \sum_{i=0}^{j_0} \sum_{k=0}^{K} \pi_{n''}(i, k)\\ &= \lim_{n''} P[X^{(n'')} \leqq j_0]\\ &\geqq \sum_{i=0}^{j_0} P_K(i)\\ &\geqq 1 - \epsilon.\end{aligned}$$

Thus, (*ii*) holds and eq. (30) follows.

It can be readily shown, using the above results together with standard theorems, that the delay probabilities in the nth model converge to those in the limiting model. (See Refs. 16 and 17.)

REFERENCES

1. S. Halfin, unpublished work.
2. F. R. Wallace, personal communication.
3. S. Halfin, "An Approximate Method for Calculating Delays for a Family of Cyclic Type Queues," B.S.T.J., this issue, pp. 1733–1754.
4. H. Kushner, *Introduction to Stochastic Control*, New York: Holt, Rinehart, and Winston, 1971, p. 211.
5. W. S. Hayward, Jr., "Traffic Engineering and Administration of Line Concentrators," Congress Record, Paper No. 23, Second International Teletraffic Congress, The Hague, 1958.
6. J. G. Kappel, personal communication.
7. C. Clos and R. I. Wilkinson, "Dialing Habits of Telephone Customers," B.S.T.J., *31*, No. 1 (January 1952), p. 46, last paragraph.
8. J. Riordan, *Stochastic Service Systems*, New York: John Wiley, 1962, pp. 109–112.
9. Ref. 7, pp. 32–67.
10. S. Karlin, *A First Course in Stochastic Processes,* New York: Academic Press, 1968, p. 194.
11. R. I. Wilkinson, "Theories for Toll Traffic Engineering in the USA," B.S.T.J., *35*, No. 2 (March 1956), pp. 421–514.
12. Ref. 11, p. 453.
13. Ref. 11, p. 454, eq. (19).
14. Ref. 10, theorem 3.2.
15. Ref. 10, pp. 206–208.
16. P. Billingsley, *Convergence of Probability Measures,* New York: John Wiley, 1968, p. 224.
17. H. Royden, *Real Analysis,* 2nd Edition, New York: MacMillan, 1968, p. 232.