THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 51, No. 6, July-August, 1972
Printed in U.S.A.

The Camp-On Problem for Multiple-Address Traffic

By ERIC WOLMAN

(Manuscript received March 3, 1972)

A communication system for multiple-address messages is described, in which a message waits in parallel queues until it can be transmitted simultaneously to all the addressed receivers. An idealized mathematical model of this system leads to a nonlinear integral equation for the stationary distribution of delays in receiver queues. A phase-plane analysis shows this equation to have a one-parameter family of solutions, one member of which is found to be the unique limiting distribution of receiver delays. Even though service times (message lengths) are not bounded, the receiver queues in this model can operate in the steady state at *critical load. Under these conditions, the probability that a server is idle is positive; and all moments of the delay distribution are finite. Computation of the delay distribution is discussed, numerical examples are given, and the behavior of the transmitter queues is analyzed.*

Predictions of this model are compared with performance parameters of simulated systems. The model is shown to be very accurate up to its critical load. For higher loads, performance depends strongly upon the number of receivers in the system. The model's discontinuity in receiver occupancy is not physically realizable, but is approached asymptotically as the number of receivers tends to infinity.

Contents

I. INTRODUCTION

The frequency of conference calls in voice telephony is very low; but many of the messages carried by some data communication systems are directed from a single transmitter to two or more destinations. In the mathematical analysis of some types of data traffic, it is the presence of these multiple-address (MA) messages that raises problems essentially different from those encountered in the classical congestion theory of telephone systems.

Methods of coping with MA traffic fall into three classes. One involves *message switching* (also called *storage*, or the store-and-forward method), in which a transmitter sends a message to a switching center or other central location at which it is stored. Copies of the message are then retransmitted more or less independently to the desired receivers. A second class makes use of *selective calling*. In the simplest example, all stations are connected to a single channel, which may be thought of as a loop without a central switch; a message on this channel may be directed to any or all of the receivers. The number of simultaneous transmissions that is possible in a system of this class is limited by the number of space-, time-, or frequency-division channels in the loop. The third class of methods is based on *line switching*, in which a switching center merely establishes connections between the terminals of a system instead of providing storage for messages in transit. Line switching for MA traffic has itself two extreme forms, *sequential* and *simultaneous*, between which lie many other types of line switching. The sequential technique requires the transmitter to send each copy of a message to the proper receiver as a separate transmission, so that the transmitter sees an MA message as a sequence of single-address messages. With simultaneous line-switching, all copies of a message are sent at once after the transmitter has been connected to all the addressed receivers.

The term "camp-on" refers to one natural implementation of simultaneous line-switching. Suppose that a transmitter serves messages

* A nonmathematical statement of the problem and of the results obtained may be found in the starred sections.

offered to the system in order of arrival, and that a queue of delayed messages can form at each transmitter. When a message reaches the head of a transmitter queue, it gives its set of addresses to some control device, thereby requesting connection to several receivers. The simplest plan is for such requests to be entered into first-come, first-served queues corresponding to individual receivers. When a message (or copy thereof) reaches the head of a receiver queue, it (i.e., other copies) may still be awaiting other receivers that are not yet idle. Then the message *camps on* the available receiving line, so that the receiver, although idle, appears busy to other traffic. Thus it is possible for a delayed message to be waiting for two receivers neither of which is actually engaged in receiving a message; but reshuffling of the order in which messages are handled, so as to avoid this situation, would require involved computations, not to mention precautions against indefinite postponement of some transmissions.

For systems in which the lines to terminals radiate from a switching center, we may think of message switching, camp-on, and sequential line-switching as the three basic or "pure" techniques for handling MA traffic, in the sense that other schemes are really combinations of these three. There is much to be said for and against each of them. Message switching requires expensive storage-facilities, and saddles the communication system with the responsibility for messages in its possession. Sequential line-switching leads to excessively heavy loading of transmitters; and receivers are used with corresponding inefficiency in camp-on systems. I cite these *dis*advantages in order to point out that, in many situations, a practical system must combine some features of at least two of the pure techniques. In fact, the advantages of many compromise schemes are well-known. One possibility is to use storage for all MA messages, while sending single-address traffic over line-switched connections. In another plan, described as simultaneous transmission to available destinations (STAD), the transmitter is connected to all the addressed receivers that are not occupied. When this transmission is complete, a new attempt is made to reach the remaining addresses. STAD avoids factitious loading of receivers while holding down the number of transmissions needed per message.

The invention, design, and analysis of good arrangements for handling MA traffic are very difficult.* Our understanding of this subject is limited, although approximate analysis, simulation, and field measure-

* These difficulties are greatest in connection with line switching, practical interest in which has decreased as technological changes have made other schemes more attractive.

ment have yielded considerable data. It seems clear that a first step toward an adequate theory must be a thorough comprehension of each of the basic techniques mentioned above.

The presence of storage in message-switching systems decouples the two stages of blocking (at the transmitters and at the receivers) enough to allow application of standard delay-theories to each stage. No satisfactory analysis of sequential line-switching has been published; but the structure of such systems is not of deep theoretical interest, partly because the key question is that of the order in which the addressees of a message are served. The camp-on situation is of crucial importance because of its combinatorial structure. This structure is described in the next section. Here it is enough to say that the service process in the transmitter queues depends on the way in which messages interact in the receiver queues.

This paper describes the analysis of a very simple and idealized model of a line-switching system with camp-on. The answer obtained is a description, valid in the limit of very large systems, of message delays during steady-state operation. This problem is one of a large class of problems, characterized by complex interactions between queues in parallel, which forms an important domain on the frontier of congestion theory. The analysis presented here is important for two reasons: It yields insight into one of the basic procedures for handling multiple-address traffic, and should therefore lead toward an understanding of more realistic models; and it embodies a modification, which may prove useful in similar problems, of a standard method in queuing theory. In addition, the remarkable behavior exhibited by this idealized model renders it of independent interest.

The problem treated here, and others not dissimilar, have been discussed by other authors. The most relevant paper is by Haenschke.[1] But discussion of other work is deferred until Section VII so that different approaches may be compared in detail.

II. A CAMP-ON MODEL

2.1 *The System*

We first consider a system with one switching-center, X transmitters, and R receivers. Messages are offered to the system in X independent arrival-processes, one for each transmitter; and each process is Poisson with rate α. Every message is addressed to exactly m of the receivers. This m is an integer greater than 1. The m addresses of a message are chosen at random uniformly from the R possibilities, and the addresses of any two messages are chosen independently. The unit of time is the mean length of the messages, and these lengths are independent random variables with the negative exponential distribution. By symmetry, all the receivers experience identical arrival processes. Let their mean rate be λ. If the system can operate in equilibrium, the number of copies of messages that must leave the transmitters is $m\alpha X$ per unit time. Since λR copies per unit time arrive in the receiver queues, the conservation equation in equilibrium must be $m\alpha X = \lambda R$. A useful form of this equation is

$$\frac{R}{X} = \frac{m\alpha}{\lambda} = k, \tag{1}$$

where we have introduced the parameter k to denote the ratio of the numbers of receivers to transmitters.

The previous paragraph, together with the description of camp-on given in Section I, completely specifies the traffic characteristics of a communication system. Except for its mean rate λ, we are ignorant of the arrival process at the receiver queues; but it is determined by the stated requirements. We also know nothing of the conditions for the existence of a steady state, although clearly λ and α cannot be too large. The "service time" of a message in a transmitter queue consists of all the time spent at the head of that queue before departure–that is, of the transmission time plus the time spent waiting for the last of the addressed receivers to become available. The situation is shown in Fig. 1 for the case in which $m = 2$. Each circle at the upper level represents a message in a transmitter queue, and at the lower level a copy in a receiver queue. The lines in the figure connect each message-symbol at the head of a transmitter queue to the symbols for the two corresponding copies. The head messages at transmitters I and V are being transmitted, to receivers 1 and 2 in the former case and 5 and 6 in the latter. The message at II is awaiting receiver 2 and camping

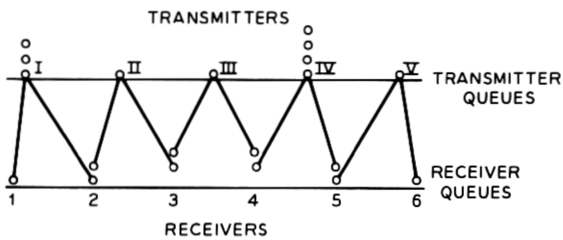

Fig. 1—The queuing discipline in a camp-on system.

on 3; and similarly for the head message at IV. Although neither of receivers 3 and 4 is receiving a message, both are being camped on and act occupied. Therefore the message at III, which announced its addresses after the head messages at II and IV did so, is, as shown in Fig. 1, effectively in third place in queuing for receivers.

The system just described is unrealistically simple in several ways. First, it consists of a single switching-center with lines to receivers and transmitters radiating from it, whereas a physical system could have many switching-centers interconnected by trunks. This difference is not critical, because we are studying principally the effect of having to wait simultaneously in m receiver-queues. Blocking due to inadequacy of trunk facilities could be made negligible in comparison. Second, signaling and switching times are taken to be zero, a mathematical convenience of long standing in queuing theory and often justified in practice. Third, each message has the same number of addresses. This assumption is far indeed from being realistic, but we shall see in Section 5.2 that it does not seriously restrict the usefulness of the model. Fourth, the transmitters and receivers are equally and independently loaded. This is a genuine restriction, especially for the receivers. A wide variation among receiver loads would represent reality better, particularly if a destination with heavy traffic could have several receivers sharing the load. Traffic flow in some physical systems is further complicated by more restrictive geometries. For example, a transmitter and receiver can be connected to a switching-center by a single line not capable of full-duplex operation (simultaneous transmission in both directions). The simple model adopted here is required to avoid prohibitive mathematical complexity, but it has the corresponding virtue of introducing no complications other than that inherent in the camp-on discipline itself.

In order to proceed, we need symbols for various portions of the time that a message spends in the system. Figure 2 shows this time-interval for a particular message, and is drawn from the viewpoint of *one* of the m receivers to which the message is addressed. At the point A in Fig. 2, the message arrives and joins a transmitter queue.

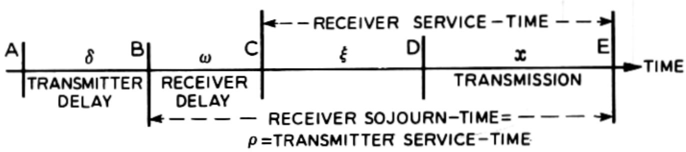

Fig. 2—Composition of the time spent by a message in the system.

At B the message reaches the head of this queue and notifies the control of its addresses; thus the message arrives in the distinguished receiver's queue (and in those of the other $m - 1$ receivers) at time B also. The interval from A to B is the *transmitter delay*, and is represented by the random variable δ.

At C, after the *receiver delay* (or waiting time) ω, the message reaches the head of the receiver queue and camps on the receiver. The longest of the m receiver-delays ends at time D, and transmission then begins. The random variable ξ represents the excess of the longest of the receiver delays over the delay at this receiver (the one for which the diagram is drawn). Of course $\xi = 0$ if the longest delay occurs at this receiver. If $\omega^{(i)}$ is the delay at the ith receiver, then ξ is defined by the equation

$$\omega^{(j)} + \xi^{(j)} = \max (\omega^{(1)}, \cdots, \omega^{(m)}) \quad \text{for each} \quad j = 1, \cdots, m.$$

The dummy indices run, of course, over the addresses of one message. Omitting superscripts for "our own" receiver, we write the simpler formula

$$\omega + \xi = \max_{i \leq m} \omega^{(i)}. \tag{2}$$

The message is actually transmitted during the interval of length x, with density function $\exp(-x)$, that runs from D to E. At time E the message ends and departs from the system.

The interval from B to E, of random length $\rho = \omega + \xi + x$, is the *sojourn time* of the message in the receiver queue. During this entire time the message occupies its transmitter, so that ρ is also the *service time* in the transmitter queue. Indeed, the transmitter queue is a single-server queue, with Poisson arrivals, whose service times ρ have an unknown distribution and are not independent.

The reader who is interested in the results of this investigation, rather than in their derivation, may now skip to Section VII, which contains a summary of the argument and a discussion of its conclusions.

2.2 *Passage to a Limit*

The trick that makes this model amenable to analysis is this: We let R and X both go to infinity while holding their ratio constant. In fact–see eq. (1)–we keep k, α, λ, and m all fixed as the system gets larger. In general terms, the purpose of this trick is to decouple the queues that operate in parallel, while preserving the essential interaction caused by each message's having to wait in m queues at once. The detailed consequences of this procedure are three in number, all of them necessary to further progress. The first of these is that the receiver queues become independent. In other words, in the infinitely large system obtained by letting R and X approach infinity, the m copies of the random variable ω that are encountered by each message are independent. I have no rigorous proof of this fact, which is, however, buttressed by the following heuristic argument.

Consider the m receivers addressed in a particular message. Before this message reaches the head of a transmitter queue, the m receiver-queues may have various lengths. Then, at time B in the sketch of Fig. 2, the message arrives simultaneously in all m queues. When, at time E, the message departs from the system, it leaves behind it the same *expected* number of messages in each of the m queues: For this message had the same sojourn-time in each of these queues, which are all subjected to arrival processes of the same intensity λ. Let us then call the departure of this message a *synchronizing event* (*S-event*) for these m queues, using "synchronizing" in a rather weak sense.

Each receiver queue participates in λ S-events per unit time, on the average, if the system is in equilibrium. Consider a fixed pair of receivers which have just participated in the same S-event. The probability that any single subsequent message addressed to receiver 1 will generate an S-event involving receiver 2 is $(m - 1)/(R - 1)$, since the $m - 1$ addresses of such a message (other than receiver 1) are chosen with equal probability from among the $R - 1$ other receivers. The S-events involving receiver 1 form a sequence of Bernoulli trials with respect to the probability of involving receiver 2, because the addresses of different messages are chosen independently. Therefore the mean interval between successive S-events involving two particular receivers is $(R - 1)/(m - 1)$ events (corresponding to one of the receivers), and this is equal to $(R - 1)/[\lambda(m - 1)]$ units of time. This quantity increases without bound when we let $R \to \infty$ while holding λ and m constant. In other words, in the infinitely large system obtained by means of our trick, the mean time between successive "synchronizations" of two given receiver queues is infinite.

Indirect methods of "synchronization" also affect the receiver queues. If receivers 1 and 3 participate in an S-event, and if shortly thereafter receivers 2 and 3 do so, we might say that the queues at receivers 1 and 2 are connected by "an S-chain of length 2." The effect of such chains is, of course, to increase the degree of statistical dependence between the states of the queues involved. But the longer an S-chain is, the less effective it can be in "synchronizing" two queues, because of the time-lag between the S-events that form the ends of the chain. Here we do not investigate the relative efficacy, in increasing the dependence of the states of any pair of receiver queues, of S-chains of different lengths. But the time required, on the average, to complete an S-chain of given length between two specified queues clearly approaches infinity as $R \to \infty$, so that the efficacy of any such chain vanishes in the infinite model.

This argument for the independence of receiver queues can be summarized by saying that as $R \to \infty$ the frequency of occurrence, of events through which the state of one receiver queue can influence the state of any other particular receiver queue, goes to zero. Also, each receiver in this statement can be replaced by any fixed, finite set of receivers.

We come now to the second consequence of passing to an infinitely large system, which is that the service times of successive messages in any transmitter queue are independent. In the finite system, the probability that the sets of m addresses of two messages have one or more receivers in common is easily found to be $m^2/R + O(R^{-2})$. This quantity tends to zero as $R \to \infty$. A similar calculation shows that, in the infinite system, the address-sets of any finite group of messages overlap with probability zero. The service times of any set of messages in a transmitter queue, which are just their respective sojourn-times in the receiver queues, therefore depend on the states of nonoverlapping (with probability one) sets of receiver queues; and these we have already found to be independent. Thus each transmitter queue is of type M/G/1–that is, has Poisson arrivals, one server, and general independent service-times.

(Notice that both these arguments are not uniform in λ: Fixing R, we can choose λ large enough to ensure a significant degree of dependence between the states of various receiver queues and also between successive transmitter service-times. Thus we must allow R to go to infinity before varying λ.)

The third consequence of our trick is that the arrival process at each receiver is Poisson. In the finite system each such process, or, in Khinchin's[2] words, stream of events, is the union of X substreams, each of intensity λ/X or $m\alpha/R$. Each event in one substream (except for an arrival at an idle transmitter) coincides with the departure of a message from a transmitter queue, for that is the instant at which the next message, if any, announces its addresses and joins m receiver-queues. Such a substream is obtained by selecting at random a fraction m/R of the events in the departure process from one transmitter, this process being itself a stream of intensity α. The resulting substreams which make up the arrival process at one receiver have, in the steady state, these properties: They are orderly and stationary (see Khinchin[2]), and, by the arguments given above, they approach mutual independence as R and $X \to \infty$. A limit theorem (see Ref. 2, Chap. 5, and the similar but slightly earlier results of Cox and Smith[3]) suggests that as $X \to \infty$ the arrival stream becomes Poisson. This result requires that one technical condition [Ref. 2, p. 50, condition (2)] be fulfilled; this is easily verified. It also requires that Khinchin's version of the theorem hold when the substreams of decreasing intensity λ/X are not independent, but merely *approach* independence as $X \to \infty$; and this is assumed here.

Let us review the properties of our final model, focusing, as in Fig. 2, on what happens to a single message. The queuing system consists of two stages. The first stage is an ordinary M/G/1 queue with an unknown distribution of service times. Service in this queue consists of a sojourn in the second-stage queue. The latter has Poisson arrivals and an exponential departure-mechanism (known as transmission) with unit mean. But transmission in the second queue begins only after the expiration of the longest of m independent, identically distributed intervals, each of which is what would ordinarily be called the waiting time in the second queue.

We constructed this model by starting with a very simple but physically realizable finite system and modifying it in an appropriate way. The properties of the final model were deduced informally. But this model is of interest in its own right, and could have been proposed for study at the start. Its properties are derived here in order to demonstrate its connection with the camp-on problem. It is also fair to say that the interactions which it is the purpose of our limiting-procedure to remove–namely, dependence of receiver delays, dependence of transmitter service-times, and deviation from Poisson character of receiver arrivals–are demonstrably small in a large finite system at moderate loads. (A system is large in this sense, if a receiver gets hardly any of its messages from any single transmitter and every message is addressed to a very small fraction of all receivers.) The effects of such interactions are discussed further in Section 5.4.

2.3 *Equation for Delay Distribution in Equilibrium*

If we knew the distribution of sojourn times in the receiver queues, which are identical with the service times in the transmitter queues, we could in principle determine the distribution of transmitter delays from the well-known formula of Pollaczek and Khinchin [Khinchin,[2]

p. 116; Cox and Smith,[4] Section 2.6, eq. (28)]. Lacking closed forms for the various quantities involved, we could still find the mean transmitter-delay, given the mean and variance of the receiver sojourn times [Khinchin,[2] p. 117; Cox and Smith,[4] Section 2.6, eq. (22)]. We therefore turn our attention exclusively to the receiver queues, which can be studied without reference to the larger system of which they are a part. The results of this investigation are then used to describe the transmitter queues, which are not mentioned again until Section VI.

We now reproduce in Fig. 3 the relevant part of Fig. 2, drawn as before for one receiver, but extended so as to show the relation between two successive messages addressed to that receiver. Subscripts refer to messages, in order of arrival; in particular, ω_n is the delay suffered by the nth message and ω_{n+1} that of the next. We also need symbols for the as-yet-unknown distribution functions for the receiver delays ω and sojourn times ρ. These are respectively F and G. Thus G_n is the sojourn-time distribution for the nth message, and F_{n+1} the delay distribution for its successor.

Proceeding in the spirit of Lindley[5] (or see Cox and Smith,[4] Section 5.3), we now relate the delay distributions of successive customers. We can do this by using two simple integral-relations, each of which amounts merely to a definition. Note first that ρ_n is the sum of two random variables: x_n, whose density function is the unit exponential, and $\omega_n + \xi_n$, which we know from (2) to be the maximum of m independent copies of ω_n. The probability that this maximum does not exceed t is the product of the probabilities that each of the m variables ω_n does not exceed t. Thus the distribution function for $\omega_n + \xi_n$ is

$$\Pr\{\omega_n + \xi_n \leqq t\} = [F_n(t)]^m.$$

The usual convolution-formula gives for $G_n(t)$, the probability that ρ_n does not exceed t, the value

$$G_n(t) = \int_0^t e^{-(t-u)} F_n^m(u)\, du, \tag{3}$$

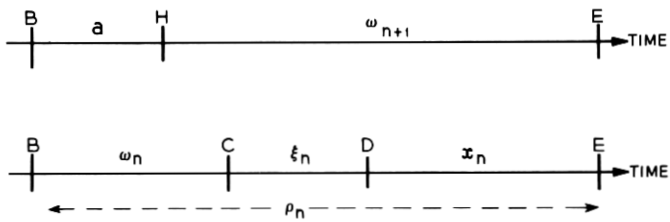

Fig. 3—Receiver delays of successive messages.

where $F_n^m(u) \equiv [F_n(u)]^m$.

An equally simple calculation yields F_{n+1}. Let a in Fig. 3 be the interval between the arrival time of the nth message (at B) and that of the next (at H). The delay ω_{n+1} runs from time H to time E, stopping then because the nth message departs. (Message $n + 1$ can be transmitted only after a further interval of length ξ_{n+1}, which may of course be zero.) Certainly ω_{n+1} is less than t if ρ_n is less than t, since time H cannot precede time B. Suppose, on the other hand, that ρ_n has the value $u > t$. Then ω_{n+1} is less than t if and only if a is greater than $u - t$. But a is an interarrival time in a Poisson process of intensity λ. It exceeds $u - t$ if and only if no arrival occurs during an interval of length $u - t$, and this event has probability $\exp(-\lambda(u - t))$. Rewriting this argument in symbols, we get the formula

$$F_{n+1}(t) = G_n(t) + \int_t^\infty e^{-\lambda(u-t)}\, dG_n(u). \tag{4a}$$

Integration by parts gives us the useful equivalent formula

$$F_{n+1}(t) = \lambda \int_t^\infty e^{-\lambda(u-t)} G_n(u)\, du. \tag{4b}$$

We can represent F_{n+1} directly in terms of F_n by eliminating G_n between eqs. (3) and (4). We note first from (3) that

$$G_n(0) = 0. \tag{5}$$

This says that no message has a sojourn time of length zero; this is to be expected because ρ_n is at least as long as x_n, whose distribution has no mass at the origin. Equation (5) also removes any doubts about the integration by parts that yields (4b), in case $t = 0$.

Equation (3) also tells us that G_n has a derivative. Differentiating the right-hand member explicitly, and denoting time-differentiation by a dot, we find that

$$\dot{G}_n(t) = -G_n(t) + F_n^m(t). \tag{6}$$

Writing $dG_n(u) \equiv \dot{G}_n(u)\, du$, we can substitute (3) and (6) into (4a). Using (4b) to eliminate the remaining appearance of G_n, we obtain the important equation

$$F_{n+1}(t) = \frac{\lambda}{1+\lambda}\left[\int_0^t e^{-(t-u)} F_n^m(u)\, du + \int_t^\infty e^{-\lambda(u-t)} F_n^m(u)\, du\right]. \tag{7}$$

A useful special case of (7) comes from setting $t = 0$. This is

$$F_{n+1}(0) = \frac{\lambda}{1+\lambda} \int_0^\infty e^{-\lambda u} F_n^m(u)\, du. \tag{8}$$

If we multiply both members of this equation by exp (λt) and divide the interval of integration into two parts at t, we find that

$$F_{n+1}(0)e^{\lambda t} = \frac{\lambda}{1+\lambda} \int_0^t e^{-\lambda(u-t)} F_n^m(u)\, du + \frac{\lambda}{1+\lambda} \int_t^\infty e^{-\lambda(u-t)} F_n^m(u)\, du.$$

We can use this to eliminate the integral from t to ∞ in (7), obtaining the "retrospective" formula

$$F_{n+1}(t) = F_{n+1}(0)e^{\lambda t} - \frac{\lambda}{1+\lambda} \int_0^t [e^{\lambda(t-u)} - e^{-(t-u)}] F_n^m(u)\, du. \tag{9}$$

It is possible to base the analysis of this problem entirely on the equations obtained so far, but much of the argument is simpler if we use an equivalent formulation which involves differential instead of integral equations. Differentiation of eq. (4b) shows that $\dot{F}_{n+1}$ exists for $t > 0$ and is given by

$$\dot{F}_{n+1}(t) = \lambda^2 \int_t^\infty e^{-\lambda(u-t)} G_n(u)\, du - \lambda G_n(t).$$

For $t = 0$ this must be interpreted as a derivative on the right; the jump in F_{n+1} at $0-$ can be ignored if we remember that $F_{n+1}(0) > 0$. Substitution of (4b) and suppression of the argument t yield

$$\dot{F}_{n+1} = \lambda F_{n+1} - \lambda G_n\, . \tag{10}$$

Since both terms of the right member have derivatives, it is also true that

$$\ddot{F}_{n+1} = \lambda \dot{F}_{n+1} - \lambda \dot{G}_n\, ,$$

where $\ddot{F}_{n+1}$ is also a right-derivative at $t = 0$. The sum of the last two equations is

$$\ddot{F}_{n+1} + (1-\lambda)\dot{F}_{n+1} - \lambda F_{n+1} = -\lambda(\dot{G}_n + G_n).$$

Using (6) to eliminate G_n and its derivative, we obtain

$$\ddot{F}_{n+1} + (1-\lambda)\dot{F}_{n+1} - \lambda F_{n+1} = -\lambda F_n^m, \tag{11a}$$

the differential version of (7). This must be accompanied by the boundary condition

$$\dot{F}_{n+1}(0) = \lambda F_{n+1}(0), \tag{11b}$$

which comes from (10) with the aid of (5). Also, since F_n is a distribution function, $0 \leqq F_n(t) \leqq 1$ for all $t \geqq 0$. Substitution of 1 as an upper bound for F_n in (7), where both kernels are positive, easily yields the fact that

$$0 \leqq F_{n+1}(t) < 1 \quad \text{for all } t \geqq 0. \tag{11c}$$

This inductive result, based essentially on the assumption that F_1 is a distribution function, serves as a second boundary-condition for (11a), one of whose homogeneous solutions is exp (λt).

The derivations of (7) and (11) given above are chosen for brevity, and both make use of equations involving G_n. The discussion that follows rests on the fact that, as self-contained descriptions of the sequence $\{F_n\}$, (7) and (11) are equivalent. We should therefore verify that each can be obtained from the other. We got (11c) from (7) anyway, and explicit differentiation of (7) quickly yields (11b) and (11a). On the other hand, treatment of (11a) as an inhomogeneous linear equation with driving function $-\lambda F_n^m$ allows us to use the standard formula for its general solution. Application of (11b) then yields the representation (9), which is the natural form when integrating in the positive t-direction. As it stands, equation (9) seems to contain the wrong homogeneous solution, but actually the integral term is a negative correction which offsets the growth of exp (λt). We can rearrange (9) so as to represent $F_{n+1}(t)$ as the sum of two quantities: One is the right-hand member of (7), which we know is bounded when $0 \leqq F_n \leqq 1$, and the other is the correction

$$e^{\lambda t}\left[F_{n+1}(0) - \frac{\lambda}{1+\lambda}\int_0^\infty e^{-\lambda u}F_n^m(u)\,du\right].$$

Since the bracketed quantity is independent of t, either it vanishes or $|F_{n+1}(t)|$ grows without limit as $t \to \infty$. The condition (11c) selects the former possibility, proving that both (8) and (7) hold. Thus the systems (7) and (11) are indeed equivalent.

Again following Lindley,[5] we impose the condition of equilibrium by specifying that two successive messages must have the same delay-distribution. The distinction between the nth message and its successor can be removed by omitting the subscripts in (7). This leaves the formula

$$F(t) = \frac{\lambda}{1+\lambda}\left[\int_0^t e^{-(t-u)}F^m(u)\,du + \int_t^\infty e^{-\lambda(u-t)}F^m(u)\,du\right]. \tag{12a}$$

This is our fundamental nonlinear integral equation for F, the equilibrium distribution of receiver delays. (Thus the distribution function for $\omega + \xi$ is F^m.) Since (12a), unlike (7), makes no implicit reference to the initial distribution F_1, we must adjoin to it the explicit condition

$$0 \leqq F(t) \leqq 1. \tag{12b}$$

III. SOLUTIONS

3.1 *Reduction to a First-Order System*

We are looking for distribution functions F that satisfy our basic system of equations (12) on the interval $0 \leqq t < \infty$. We expect on physical grounds that (12) has solutions for sufficiently small positive λ. We also expect that, for large enough values of λ, (12) has no solutions; for these large arrival-rates the queue lengths and delays should increase without limit, and no steady state should exist.

In this section we show that indeed there exists a critical arrival-rate Λ above which (12) has no solutions, and that for $\lambda \leqq \Lambda$ it has infinitely many solutions. In the latter case the solutions fall into two groups, one of which contains a single distribution and the other a one-parameter family of distributions. One's natural inclination to choose the distinguished solution as the answer to the physical problem encounters a major difficulty: The set of distinguished solutions obtained by varying λ has the property that their initial values, which represent the respective probabilities of finding a receiver idle, are bounded away from zero as λ approaches Λ from below. The paradoxical implication that the receivers have spare capacity at their critical load makes it unclear whether the idealized problem (with $R = \infty$) has a unique solution at all. This issue is discussed further below and resolved in Section IV.

We begin by observing two consequences of the system (12). First, its solutions must be what we may call *unlimited*. According to (12b), the right member of (12a), evaluated for $F = 1 = F^m$, is an upper bound for F; explicitly,

$$\frac{\lambda}{1+\lambda}\left[e^{-t}\int_0^t e^u\,du + e^{\lambda t}\int_t^\infty e^{-\lambda u}\,du\right] = 1 - \frac{\lambda}{1+\lambda}e^{-t}. \tag{13}$$

[This is the formula alluded to in the derivation of eq. (11c).] This result rules out functions F (called *limited*) which reach 1 at finite values of t. They would correspond, if they existed, to operation with receiver delays bounded above by some finite limit; and this situation seems physically unreasonable because transmission times are not bounded. In particular, F cannot reach the value 1 with positive slope.

Improper equilibrium-distributions, corresponding to a positive probability of infinite delay, also cannot occur. This situation is represented by distributions which never increase beyond some positive value less than 1. That this cannot happen is shown next; but here we note in passing the trivial exception that $F \equiv 0$ is a formal solution of (12).

By defining the kernel

$$k_\lambda(t, u) \equiv \begin{cases} [\lambda/(1+\lambda)] \exp(-(t-u)), & u \leqq t, \\ [\lambda/(1+\lambda)] \exp(-\lambda(u-t)), & u \geqq t, \end{cases} \tag{14}$$

we can write (12a) in the form

$$F(t) = \int_0^\infty k_\lambda(t, u) F^m(u)\, du. \tag{15}$$

Suppose there were a bound $b < 1$ such that $0 \leqq F \leqq b$ for some solution F of (12a) and (15). Then, k_λ being positive,

$$F(t) \leqq b^m \int_0^\infty k_\lambda(t, u)\, du < b^m,$$

with the final inequality coming from (13). Substitution of this result into (15) shows that $F(t) < b^{m^2}$; thus $F(t) < b^{m^j}$ for arbitrarily large j by iteration, and so $F(t) \equiv 0$. Therefore all nontrivial solutions have the property that

$$F(t) \to 1 \quad \text{as} \quad t \to \infty \tag{16}$$

and are called *proper*.

One other useful formula comes directly from (12a). Differentiation of that equation tells us that

$$\dot{F}(t) = \lambda F(t) - \lambda \int_0^t e^{-(t-u)} F^m(u)\, du.$$

Since the last term is nonpositive,

$$\dot{F}(t) \leqq \lambda F(t), \quad t \geqq 0. \tag{17}$$

It is now convenient to switch to the differential formulation of the problem. Omitting all subscripts from (11), we get for equilibrium the system

$$\ddot{F} + (1-\lambda)\dot{F} = \lambda F - \lambda F^m, \tag{18a}$$

$$\dot{F}(0) = \lambda F(0), \tag{18b}$$

$$0 \leqq F(t) < 1 \quad \text{for} \quad 0 \leqq t < \infty, \tag{18c}$$

which can also be derived from (12) and the fact that solutions of (12) are unlimited. Since t does not appear explicitly in (18a), we can transform it into an equation of first order by introducing a function $p(F)$ with these properties:

$$p(F) = \dot{F}; \qquad \ddot{F} = \frac{d\dot{F}}{dt} = \frac{dp}{dt} = \frac{dp}{dF}\cdot\frac{dF}{dt} = p'p, \tag{19}$$

where $p' \equiv dp/dF$. We also use the notation $F_0 \equiv F(0)$ so that the integration for $p(F)$ can begin with $p(F_0)$; the previous subscripts should cause no confusion, since there is no zeroth message. Equations (18a) and (19) yield, after some rearrangement,

$$p' = \frac{\lambda(F - F^m)}{p} - (1 - \lambda). \tag{20a}$$

The boundary condition (18b) becomes

$$p(F_0) = \lambda F_0 . \tag{20b}$$

Admissible solutions of (20a) must have three other properties. Since we are seeking only solutions corresponding to distribution functions, which must be monotone, p must be nonnegative. Second, (17) holds. We also have the condition (18c), except that, since every solution of (18) satisfies (16) but is unlimited, every solution of (20a) must include the point $(F, p) = (1, 0)$ (though with parameter $t = \infty$). In symbols,

$$0 \leqq p \leqq \lambda F; 0 \leqq F < 1 \quad \text{except at} \quad (F, p) = (1, 0). \tag{20c}$$

Our original problem is now reduced to that of solving the first-order system (20). One more integration then yields $F(t)$, as we see from the first part of (19), which can be written $dt = dF/p(F)$.

3.2 *Trajectories in the Phase Plane*

Equation (20a) is best studied in the Fp-plane, which is properly called the phase plane for (18) because $p = \dot{F}$. Figure 4 shows the relevant region of this plane for the case in which $\lambda < \Lambda$. According to (20c), meaningful solutions must lie above the F-axis, within the strip $0 \leqq F \leqq 1$, and below the oblique line $p = \lambda F$. The formula (20a) defines a vector field, and we must study those of its integral curves that lie within the triangle just described. These integral curves, or trajectories, are parametrized by t according to the results of the

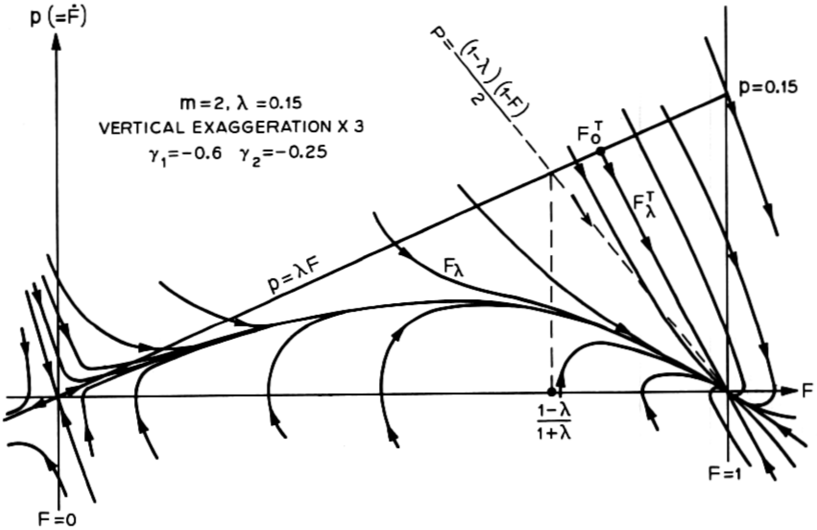

Fig. 4—Phase-plane trajectories for subcritical load.

second integration. The boundary condition (20b) tells us that the desired trajectories begin (with $t = 0$ and $F = F_0$) on the line $p = \lambda F$.

What does the vector field look like? Since F is nondecreasing, the field points to the right (for increasing t), or vertically if $p = 0$. Along the right edge of the triangle, where $F = 1$ and $p > 0$, (20a) shows that $p' = -(1 - \lambda)$. Some of these field vectors are shown in Fig. 4. They are drawn for $\lambda < 1$; this inequality is shown below to be appropriate. Along the top edge of the triangle, (20a) reduces to $p' = \lambda - F^{m-1}$. This agrees with the right-edge field at the corner $(1, \lambda)$, and points increasingly upward as we progress toward the origin. The field points into the triangle everywhere on the open top edge; in the limit it is parallel to the top edge at (0, 0). Along the bottom edge, $p' = +\infty$. The plus sign is chosen because, for fixed F and small positive p, p' assumes large positive values. The field is singular at the two corners (0, 0) and (1, 0), since p' assumes different limiting values depending on the direction from which each of these points is approached.

In studying these two singularities we mainly follow Chapter 15, "Perturbation Theory of Two-Dimensional Real Autonomous Systems," of Coddington and Levinson;[6] however, the present discussion is self-contained, though succinct. We begin by writing (20a) as a pair of simultaneous parametric equations. The standard form, obtained from (19) and (20a), is

$$\dot{F} = p, \qquad \dot{p} = \lambda F - (1 - \lambda)p - \lambda F^m. \tag{21}$$

The system (21) is suitable for studying the singularity at (0, 0), but it is convenient to examine the one at (1, 0) first. For this purpose we introduce the variable $\phi = 1 - F$, which transforms (21) into

$$\dot{\phi} = -p, \qquad \dot{p} = \lambda(m - 1)\phi - (1 - \lambda)p - \lambda \sum_{j=2}^{\infty} \binom{m}{j}(-\phi)^j \tag{22}$$

in the ϕp-plane. (Of course the binomial expansion terminates when m is an integer. The significance of nonintegral m is discussed in Section 5.2.) Omission of the higher-order terms leaves the unperturbed system

$$\dot{\phi} = -p, \qquad \dot{p} = \lambda(m - 1)\phi - (1 - \lambda)p, \tag{23}$$

which is linear and has its only singularity at $(\phi, p) = (0, 0)$. As on p. 371 of Ref. 6, this can be written in the form

$$\begin{pmatrix} \dot{\phi} \\ \dot{p} \end{pmatrix} = A_1 \begin{pmatrix} \phi \\ p \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 0 & -1 \\ \lambda(m - 1) & -(1 - \lambda) \end{pmatrix}. \tag{23a}$$

The determinant of A_1 is $\lambda(m - 1)$, which does not vanish because $\lambda > 0$ and $m > 1$.

The eigenvalues of A_1 satisfy the quadratic equation

$$\gamma^2 + (1 - \lambda)\gamma + \lambda(m - 1) = 0 \tag{24}$$

with discriminant $(1 - \lambda)^2 - 4\lambda(m - 1)$, and can therefore be written

$$\gamma_1 = -(1/2)\{1 - \lambda + [(1 + \lambda)^2 - 4\lambda m]^{\frac{1}{2}}\}, \qquad \gamma_2 = -(1/2)\{1 - \lambda - [(1 + \lambda)^2 - 4\lambda m]^{\frac{1}{2}}\}. \tag{25}$$

The discriminant of these expressions vanishes when $\lambda^2 - 2(2m - 1)\lambda + 1 = 0$; that is, at these two values of λ:

$$\Lambda(m) = 2m - 1 - 2(m^2 - m)^{\frac{1}{2}}, \tag{26a}$$

$$\Upsilon(m) = 2m - 1 + 2(m^2 - m)^{\frac{1}{2}}. \tag{26b}$$

(These symbols are mnemonic for "lower" and "upper"; also $\Lambda(m)$ turns out to be the critical load Λ mentioned above.) The discriminant in (25) represents a parabola in λ opening upwards, and it takes the values 1 at $\lambda = 0$ and $4(1 - m)$ at $\lambda = 1$. Since $m > 1$, we see that $0 < \Lambda(m) < 1 < \Upsilon(m)$ and that the discriminant is negative only when λ lies between Λ and Υ. When the discriminant is positive it is smaller than $(1 - \lambda)^2$, so that the eigenvalues γ_i are of one sign. The range of arrival rates λ can now be divided as follows, according to the nature of the eigenvalues of A_1:

Case 1. $0 < \lambda < \Lambda$: γ_i distinct, real, negative;
Case 2. $\lambda = \Lambda$: both $\gamma_i = (\lambda - 1)/2 < 0$;
Case 3. $\Lambda < \lambda < \Upsilon$: γ_i complex conjugate;
Case 4. $\lambda = \Upsilon$: both $\gamma_i = (\lambda - 1)/2 > 0$;
Case 5. $\lambda > \Upsilon$: γ_i distinct, real, positive.

This classification allows us to put A_1 into the canonical forms listed in Ref. 6. We treat the cases in order, beginning with the first and most important, which is represented in Fig. 4.

Case 1: We introduce new coordinates (x_1, y_1) by means of the linear transformation

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = T_1 \begin{pmatrix} \phi \\ p \end{pmatrix}. \tag{27}$$

If we choose for this operator the real nonsingular matrix

$$T_1 = \begin{bmatrix} \lambda(m - 1) & \gamma_2 \\ \lambda(m - 1) & \gamma_1 \end{bmatrix} \tag{28a}$$

with inverse

$$T_1^{-1} = \delta_1^{-1} \begin{bmatrix} \gamma_1 & -\gamma_2 \\ -\lambda(m - 1) & \lambda(m - 1) \end{bmatrix}, \tag{29a}$$

where the determinant $\delta_1 < 0$, then the linear system in (23a) becomes

$$\begin{pmatrix} \dot{x}_1 \\ \dot{y}_1 \end{pmatrix} = J_1 \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \qquad J_1 = T_1 A_1 T_1^{-1} = \begin{bmatrix} \gamma_2 & 0 \\ 0 & \gamma_1 \end{bmatrix}. \tag{30a}$$

The transformed matrix J_1 has the canonical form listed as (II) on p. 371 of Ref. 6; the trajectories of (30a) near the x_1y_1-origin (an improper node) are sketched in Fig. 5 on p. 373 of Ref. 6. All these trajectories approach $(x_1, y_1) = (0, 0)$ as $t \to \infty$. The positive and negative y_1-axes are trajectories; all the other integral curves have zero as their limiting slope at the origin and reach the origin tangent to the x_1-axis.

Since the nonlinear system (22) can be written

$$\begin{bmatrix} \dot{\phi} \\ \dot{p} \end{bmatrix} = A_1 \begin{bmatrix} \phi \\ p \end{bmatrix} + \begin{bmatrix} 0 \\ -\lambda \sum_{j=2}^{\infty} \binom{m}{j} (-\phi)^j \end{bmatrix},$$

the corresponding transformed system is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{y}_1 \end{bmatrix} = J_1 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + T_1 \begin{bmatrix} 0 \\ -\lambda \sum_{j=2}^{\infty} \binom{m}{j} (-\delta_1)^{-j} (\gamma_1 x_1 - \gamma_2 y_1)^j \end{bmatrix},$$

in which, by virtue of (27) and (29a), the expression $\delta_1^{-1}(\gamma_1 x_1 - \gamma_2 y_1)$ has been substituted for ϕ. Thus the canonical form of the nonlinear system is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{y}_1 \end{bmatrix} = J_1 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} -\lambda\gamma_2 \Sigma_2^{\infty} \binom{m}{j} (-\delta_1)^{-j} (\gamma_1 x_1 - \gamma_2 y_1)^j \\ -\lambda\gamma_1 \Sigma_2^{\infty} \binom{m}{j} (-\delta_1)^{-j} (\gamma_1 x_1 - \gamma_2 y_1)^j \end{bmatrix}. \tag{31a}$$

If we call the perturbing vector in this equation (f_1, f_2), we can apply Theorem 5.1 on p. 384 of Ref. 6 by establishing two properties of the f_i. Let $r^2 = x_1^2 + y_1^2$. Then f_1 and f_2 must be $o(r)$ as $r \to 0$; and also $\partial f_1/\partial x_1$ and $\partial f_2/\partial x_1$ must exist and be continuous in a neighborhood of the $x_1 y_1$-origin. It is easy to see from the expressions in (31a) that the perturbing functions f_i satisfy these hypotheses. The theorem cited then says that the trajectories of (31a) have the same topological behavior near the $x_1 y_1$-origin as those of (30a); namely, that all trajectories near the origin approach the origin as $t \to \infty$: one each becoming tangent to the positive and negative y_1-axes at the origin, and all the rest with limiting slope zero.

This improper node in the $x_1 y_1$-plane can be transplanted back to the Fp-plane through application of T_1^{-1} and then of the substitution $F = 1 - \phi$. From (27) and (29a), and because (24) shows that $\gamma_1\gamma_2 = \lambda(m-1)$, the negative y_1-axis has direction $(1, -\gamma_1)$ in the first quadrant of the ϕp-plane; likewise the positive x_1-axis has the ϕp-direction $(1, -\gamma_2)$. These directions then become lines of slope γ_1 and γ_2 respectively, passing through the point $(F, p) = (1, 0)$. Thus exactly one trajectory of (21) reaches $(F, p) = (1, 0)$ with slope γ_1 from inside the triangle of Fig. 4; and infinitely many trajectories reach this singular point, from inside the triangle, with slope γ_2. (Note that $\gamma_1 < \gamma_2 < 0$.) Typical trajectories near $(F, p) = (1, 0)$ are shown in Fig. 4.

A similar analysis covers the singularity at the Fp-origin. The unperturbed form of (21) involves the matrix

$$A_0 = \begin{bmatrix} 0 & 1 \\ \lambda & -(1-\lambda) \end{bmatrix} \tag{23b}$$

with determinant $-\lambda < 0$. The eigenvalues are -1 and λ–that is, real and of opposite sign for all relevant values of λ. The real nonsingular matrix

$$T_0 = \begin{bmatrix} \lambda & -1 \\ 1 & 1 \end{bmatrix} \tag{28b}$$

with inverse

$$T_0^{-1} = (1+\lambda)^{-1} \begin{bmatrix} 1 & 1 \\ -1 & \lambda \end{bmatrix} \tag{29b}$$

brings in new coordinates (x_0, y_0) as in (27), and puts A_0 in the canonical form

$$J_0 = T_0 A_0 T_0^{-1} = \begin{bmatrix} -1 & 0 \\ 0 & \lambda \end{bmatrix} \tag{30b}$$

listed as (IV) on p. 371 of Ref. 6. The corresponding singularity is a saddle point, as sketched in Fig. 9 on p. 374 of Ref. 6. The nonlinear system (21) transforms into

$$\begin{bmatrix} \dot{x}_0 \\ \dot{y}_0 \end{bmatrix} = J_0 \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} + \lambda \begin{bmatrix} [(x_0 + y_0)/(1+\lambda)]^m \\ -[(x_0 + y_0)/(1+\lambda)]^m \end{bmatrix}. \tag{31b}$$

Calling the perturbing functions f_1 and f_2 again, we invoke part (a) of Theorem 6.1 on p. 387 of Ref. 6. The required hypothesis is that both f_i be $o(r)$ (with $r^2 = x_0^2 + y_0^2$), which is true because $m > 1$. Then the trajectories of (31b) behave as follows: One pair of integral curves leaves the origin tangent to the y_0-axis; *at least* one pair approaches the origin along the x_0-axis; and the other nearby trajectories resemble modified hyperbolae. (Whether the pair reaching the origin tangent to the x_0-axis is unique or not, depends on m; but this question need not concern us.) From (29b), the direction of the positive y_0-axis is $(1, \lambda)$ in the Fp-plane–that is, along the upper edge of the triangle of Fig. 4. The x_0-axis has Fp-direction $(1, -1)$. Thus the field near the Fp-origin has integral curves as shown in Fig. 4.

To support the remaining features of Fig. 4, we need only show that the trajectories passing into the triangle through some (leftmost) portion of its oblique upper edge do in fact reach the point (1, 0). [The contrary possibility would be that all such trajectories pass out through the right edge of the triangle, while the family of trajectories that have been shown to reach (1, 0) are among those that enter the triangle vertically through its bottom edge.] In fact all trajectories beginning to the left of $F = (1 - \lambda)/(1 + \lambda)$ have the desired property; for this point is the intersection of the upper edge $p = \lambda F$ with a straight line of slope $-(1 - \lambda)/2$ passing through (1, 0), and we now show that no trajectory can cross this line from left to right.

To prove this we must demonstrate that the vector field points more steeply downward, along this straight line, than the slope of the line itself; the field is so shown in Fig. 4. If we substitute the equation of the line in question, $p = (1 - \lambda)(1 - F)/2$, into the expression (20a) for p', we obtain $[2\lambda(F - F^m)/(1 - \lambda)(1 - F)] - (1 - \lambda)$ as the slope of the vector field along this straight line. This quantity must be more negative than the slope $-(1 - \lambda)/2$ of the straight line. The corresponding inequality can be written, after rearrangement,

$$(F - F^m)/(1 - F) < (1 - \lambda)^2/4\lambda, \quad (1 - \lambda)/(1 + \lambda) \leqq F < 1. \tag{32}$$

The left member of this inequality is an increasing function of F on $0 < F < 1$ for all real $m > 1$. (The proof is trivial for integral m, when the division can be carried out explicitly. For nonintegral m, it is perhaps easiest to show that the numerator of the derivative of this fraction is a strictly decreasing function which reaches zero at $F = 1$.) This function of F approaches $m - 1$ as $F \to 1$. Therefore (32) is satisfied if and only if

$$m - 1 \leqq (1 - \lambda)^2/4\lambda. \tag{33}$$

But this statement is true (with strong inequality) for Case 1, as the discussion on pp. 1381–2 shows.

We have shown that a trajectory can cross the line $p = (1 - \lambda)(1 - F)/2$ only *downward*. Since the slope of this line is $(\gamma_1 + \gamma_2)/2$, the qualitative aspects of Fig. 4 (which is in fact drawn to scale for $m = 2$, $\lambda = 0.15$) are verified. In particular, the unique trajectory that reaches $(F, p) = (1, 0)$ with the steeper slope γ_1 must begin, with $t = 0$ on the line $p = \lambda F$, at an abscissa F_0^T that lies to the right of the intercept $(1 - \lambda)/(1 + \lambda)$.

Before discussing the bearing of Fig. 4 on the original problem, we investigate the form of the integral curves of (21) for the remaining ranges of λ.

Case 2: When $\lambda = \Lambda$ and $\gamma_1 = \gamma_2 = \gamma < 0$, the trajectories near the Fp-origin are as shown in Fig. 4, since the analysis on pp. 1384–5 goes through unchanged. For the other singularity we follow the calculation of the previous case, beginning with the introduction of new coordinates (x_2, y_2) at the ϕp-origin by means of the matrix

$$T_2 = \begin{pmatrix} \gamma & 1 \\ 1 & 4/(\lambda - 1) \end{pmatrix}. \tag{28c}$$

It has determinant unity and the inverse

$$T_2^{-1} = \begin{pmatrix} 4/(\lambda - 1) & -1 \\ -1 & \gamma \end{pmatrix}, \tag{29c}$$

and so A_1 in (23a) assumes the canonical form

$$J_2 = \begin{pmatrix} \gamma & 0 \\ 1 & \gamma \end{pmatrix} \tag{30c}$$

listed as (III) on p. 371 of Ref. 6. The trajectories of the linear system are as shown in Fig. 7 on p. 373 of Ref. 6, and all reach the origin tangent to the y_2-axis in the first and third quadrants of the x_2y_2-plane. From (29c), we can substitute $\phi = (2x_2/\gamma) - y_2$ into (22) and obtain the canonical form

$$\begin{pmatrix} \dot{x}_2 \\ \dot{y}_2 \end{pmatrix} = J_2 \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} + \begin{pmatrix} -\lambda \Sigma_2^\infty \binom{m}{j} (-\gamma)^{-j} (2x_2 - \gamma y_2)^j \\ -(2\lambda/\gamma) \Sigma_2^\infty \binom{m}{j} (-\gamma)^{-j} (2x_2 - \gamma y_2)^j \end{pmatrix} \tag{31c}$$

for the nonlinear system.

This system is treated in problem 10 (p. 346), Chapter 13 of Ref. 6. (No general theorem, analogous to the one used for Case 1, is given in Chapter 15. Rather than establishing a new framework suitable for importing such a theorem from another source, it seems simpler to use this weaker but adequate result.) Again calling the vector that perturbs the linear system (f_1, f_2), we must show that both the f_i have continuous first partial derivatives in a neighborhood of the x_2y_2-origin; that the f_i and their first partials vanish at the origin; and that each f_i is $O([\,|x_2| + |y_2|\,]^b)$ for some $b > 1$. These hypotheses are easily verified. Then the answer given to the problem just cited is that all trajectories in a neighborhood of the origin can be represented parametrically in the form

$$\begin{aligned} x_2(t) &= c_1 \exp(\gamma t) + O(\exp((\gamma - \epsilon)t)), \\ y_2(t) &= c_1 t \exp(\gamma t) + c_2 \exp(\gamma t) + O(\exp((\gamma - \epsilon)t)), \end{aligned} \tag{34}$$

for some $\epsilon > 0$. (The coefficients c_1 and c_2 are appropriate constants, and every choice of c_1 and c_2 corresponds to a trajectory.) Thus, in particular, all these trajectories approach the x_2y_2-origin, as $t \to \infty$, tangent to the y_2-axis.

As in Case 1, T_2^{-1} and the relation $F = 1 - \phi$ enable us to move this (degenerate) improper node back to the Fp-plane. The y_2-axis becomes a straight line of slope γ passing through $(F, p) = (1, 0)$. The trajectories differ from those shown in Fig. 4 only in that all those reaching $(1, 0)$ from inside the triangle have limiting slope γ. But again there is an uppermost trajectory, which corresponds to the y_2-axis for the unperturbed system $(\dot{x}_2, \dot{y}_2) = J_2(x_2, y_2)$ and to the choice $c_1 = 0$ in (34). This distinguished trajectory crosses the edge $p = \lambda F$ (with $t = 0$) to the right of the abscissa $(1 - \lambda)/(1 + \lambda)$ by the argument used in Case 1, since (33) holds (with equality) in Case 2. Higher trajectories pass out of the triangle through its right-hand edge.

Case 3: When λ lies between Λ and Υ, the eigenvalues of A_1 are complex conjugates. The transformed matrix J_3 assumes the canonical form (V) of Ref. 6, p. 371; except that when $\lambda = 1$, so that from (25) the real part of γ_i vanishes, form (VI) occurs. In these situations, the x_3y_3-origin is a spiral point and a center, respectively, for the unperturbed system. According to Theorems 2.2 (p. 376) and 4.1 (p. 382) of Ref. 6, the nonlinear system also has either a center or a spiral point at $(x_3, y_3) = (0, 0)$, and therefore behaves likewise at $(F, p) = (1, 0)$. In this case, the system (21) can have no trajectories which reach this point from the edge $p = \lambda F$ and stay within the triangle of Fig. 4.

Cases 4 and 5: Similar arguments cover the situation when $\lambda \geqq \Upsilon > 1$; but it is even easier to observe from (20a) that $p' > 0$ everywhere in the relevant triangle, so that each trajectory which starts with $t = 0$ on the edge $p = \lambda F$ must pass out through the right-hand edge of the triangle with an ordinate exceeding $p(F_0)$ at $F = 1$.

* * *

We saw in Section 3.1 that an admissible solution to the systems (12) and (18) corresponds to a parametrized trajectory of (20a) that starts (with $t = 0$) on the line $p = \lambda F$, proceeds to the right within the triangle shown in Fig. 4, and ends (with $t = \infty$) at the point $(F, p) = (1, 0)$. The preceding discussion of Case 1 shows that, when $\lambda < \Lambda(m)$, there are two classes of admissible solutions. The first contains a single member corresponding to the unique trajectory, shown in Fig. 4, which reaches (1, 0) with slope γ_1 and starts with abscissa F_0^T. We call this solution F_λ^T. This function is represented in the phase plane by the graph of a function $p(F)$, but the latter corresponds through another integration to a distribution $F_\lambda^T(t)$: hence the symbol. The superscript is explained below; and the function F_λ^T also depends on the suppressed parameter m.

The second class of solutions is a one-parameter family represented in the phase plane by the trajectories which reach (1, 0) with slope γ_2 and begin on $p = \lambda F$ with abscissae lying in the open interval $(0, F_0^T)$. Each member of this family is called an F_λ. The superscript T was introduced as a mnemonic for "top": Not only does F_λ^T lie on top of all the F_λ in the phase plane, but also, because $dt = dF/p(F)$, it has a smaller value of the parameter t for given F than any F_λ and therefore lies above all the F_λ in the tF-plane.

The integral curves of (20a) in the triangle of Fig. 4, including those that enter through the F-axis and those above F_λ^T that correspond to distributions that are not unlimited, constitute a partition of the interior of the triangle: That is, each point of the interior lies on exactly one integral-curve. Nevertheless, those that end at (1, 0) do so only with the slopes given by (25). An F_λ starting very close to F_0^T, for instance, lies just below F_λ^T and turns, very near (1, 0), to make the angle arctan γ_2 with the F-axis.

As shown in the discussion of Case 2, the situation is similar when $\lambda = \Lambda(m)$. The only difference is that F_λ^T and all the trajectories F_λ reach (1, 0) with the same slope $\gamma = \gamma_1 = \gamma_2$. On the other hand there are no admissible solutions when $\lambda > \Lambda$, for in Cases 3, 4, and 5 every trajectory corresponds to a distribution which is not unlimited. That is, when λ exceeds Λ, there is no equilibrium distribution of receiver delays; and this is why we call Λ the critical load.

Is it possible that F_λ^T and all the members of $\{F_\lambda\}$ are meaningful "answers" to the original problem when $\lambda \leqq \Lambda$? Certainly it is not clear how one would select the "correct" solution: On the one hand, the family $\{F_\lambda\}$ forms an open set corresponding to the open interval $(0, F_0^T)$ containing its end-points, and it is hard to see why any particular member of such a family should be "better" than its fellows. On the other hand, there is a substantial objection to the idea that

F_λ^T, the only distinguished solution of (20), should be the unique answer. For if F_λ^T is the meaningful solution, then F_0^T represents the probability of finding a receiver idle, and $(F_0^T)^m$ the probability of finding all m addressed receivers idle. Is it possible that this quantity should be positive even when $\lambda = \Lambda(m)$–that is, when even the slightest increase in load would preclude the possibility of an equilibrium distribution of receiver delays? On the contrary, innumerable other problems in queuing theory suggest that the value of F_0 corresponding to the true solution of (12) should approach zero as $\lambda \uparrow \Lambda(m)$. Yet it seems that even the idealized problem should dictate a unique solution.

IV. UNIQUENESS

We are now faced with a one-parameter family of stationary delay-distributions satisfying eqs. (12). But not every *stationary* distribution need be a *limiting* distribution; and so we return to (7) and other equations of Section 2.3 to study the sequence of delay distributions corresponding to successive messages arriving at a particular receiver queue. (Since the argument of this section is more suited to formal exposition than the preceding material, the results are given as theorems rather than discursively.)

Some additional notation is required, and we begin with a restatement of the key equation

$$F_{n+1}(t) = \frac{\lambda}{1+\lambda}\left[\int_0^t e^{-(t-u)}F_n^m(u)\,du + \int_t^\infty e^{-\lambda(u-t)}F_n^m(u)\,du\right]. \qquad (7)$$

It is convenient to define the power operator $\mathcal{P}_m$ such that

$$F^m = \mathcal{P}_m F \qquad (35a)$$

for real m and the integral operator

$$\mathfrak{M}_\lambda \equiv \int_0^\infty du\ k_\lambda(t, u) \qquad (35b)$$

where the kernel k_λ is given by (14), so that, with

$$\mathfrak{N}_\lambda \equiv \mathfrak{M}_\lambda \mathcal{P}_m\,, \qquad (35c)$$

we can write (7) in the form

$$F_{n+1} = \mathfrak{N}_\lambda F_n\,. \qquad (36)$$

(The dependence of $\mathfrak{N}_\lambda$ on m is suppressed.) All functions of t mentioned in this section have domain $[0, \infty)$.

Lemma 1: If F_n is a (nontrivial) distribution function (d.f.), then F_{n+1} is strictly increasing.

Proof: By hypothesis F_n is nondecreasing, so that $F_n^m(t) \geqq F_n^m(u)$ for $0 \leqq u \leqq t$. Thus eq. (3) shows that for all t ($\geqq 0$, of course)

$$G_n(t) \leqq F_n^m(t) \int_0^t e^{-(t-u)}\, du = F_n^m(t)(1 - e^{-t})$$
$$< F_n^m(t).$$

[We write this $G_n < F_n^m$; such inequalities between functions, written without restriction, are to be interpreted pointwise on $[0, \infty)$.] Therefore by (6) $\dot{G}_n > 0$, and so also, using (4a),

$$\int_t^\infty e^{-\lambda(u-t)}\, dG_n(u) = F_{n+1}(t) - G_n(t) > 0.$$

[Notice that this inequality holds even when, because $F_n(v) = 0$, the preceding strong inequalities fail for $u \leqq v$.] From (10) we see that $\dot{F}_{n+1} > 0$. □

The next lemma shows that F_{n+1} is unlimited–that is, assigns positive probability to arbitrarily long delays.

Lemma 2: If $F_n \leqq 1$, then $F_{n+1} < 1$.

Proof: The statement of the lemma is contained in (11c). □

Lemma 3: If $F_n \to 1$ as $t \to \infty$, then $F_{n+1}(t) \to 1$ as $t \to \infty$. (In words, F_{n+1} is proper if F_n is proper.)

Proof: We showed in proving Lemma 1 that $F_{n+1} > G_n$. Thus $G_n < 1$ by Lemma 2. If G_n were bounded away from 1, then by (6), $\dot{G}_n$ would be positive and bounded away from 0 for sufficiently large t. This is impossible; therefore $G_n \to 1$ as $t \to \infty$. Since $F_{n+1} > G_n$, $F_{n+1} \to 1$ also. □

[The fact that solutions of (12) must be proper and unlimited was discussed on pp. 1377–8.]

Theorem 1: If F_1 is a proper distribution function, then the remainder of the sequence generated by (7) and (36) consists of d.f.s which are proper, unlimited, and strictly increasing.

Proof: Apply Lemmata 1, 2, and 3 inductively on n. □

The hypothesis of Theorem 1 covers the case in which $F_1 = 1$; that is, $F_1(t) = 1\ \forall t \geqq 0$. In this case, with probability one the first message suffers no delay at the receiver. In physical terms, the first message finds the system idle. We now study the system's simplest kind of transient behavior, beginning with no messages present and letting the receivers run indefinitely at a fixed subcritical load. In the discussion that follows, we represent by $F_n[F_1, \lambda]$ the nth element of the sequence generated by (36); this notation specifies both the starting function F_1 and the parameter λ of the operator $\mathfrak{M}_\lambda$. The value of this function at t is $F_n[F_1, \lambda](t)$.

We need this preliminary result:

Lemma 4: Let F_1 be a d.f. If $F_n[F_1, \lambda] \to L$ as $n \to \infty$, then L is a solution of (12).

Proof: Because L is a pointwise limit of functions which satisfy (11c), L itself must satisfy (12b). Using (35) and (36), we can write (7) in the form*

$$F_{n+1}(t) = \int_0^\infty k_\lambda(t, u) F_n^m(u)\, du, \tag{37}$$

which corresponds to the equilibrium equation (15). For each t, the sequence $\{k_\lambda(t, u)F_n^m(u)\}$ approaches $k_\lambda(t, u)L^m(u)$ as $n \to \infty$ and is bounded by $k_\lambda(t, u)$, which is an integrable function of u according to (13). These are the hypotheses of Lebesgue's Dominated Convergence Theorem, which tells us that

$$\lim_{n\to\infty} \int_0^\infty k_\lambda(t, u) F_n^m(u)\, du = \int_0^\infty k_\lambda(t, u) L^m(u)\, du.$$

By (37) and the hypothesis of the present lemma, the left member of the last equation is $L(t)$. Thus L indeed satisfies (12a). □

We are now ready to prove

Theorem 2: Let $F_1 = 1$ and choose a fixed $\lambda \leqq \Lambda(m)$. Then the sequence $\{F_n[F_1, \lambda]\}$ approaches F_λ^T.

Proof: Setting $F_1 = 1$ in (7) shows as in (13) that $F_2(t) = 1 - [\lambda/(1 + \lambda)] \exp(-t)$, so that $F_2 < F_1$. From (35) and (36), $F_{n+1} = \mathfrak{M}_\lambda F_n^m$. Writing this also for F_n, we get by subtraction

---

* Since the difference kernel k_λ is the density function of the difference $x - a$ between a transmission time and an interarrival time, we recognize (37) as a consequence of the equation $\omega_{n+1} = \max[0, \max(\omega_n^{(1)}, \cdots, \omega_n^{(m)}) + x_n - a_n]$. This relation generalizes the familiar recurrence for a single-server queue, in which $\max \omega_n^{(i)}$ reduces to ω_n.

$$F_n - F_{n+1} = \mathfrak{M}_\lambda(F_{n-1}^m - F_n^m). \tag{38}$$

Since $\mathfrak{M}_\lambda$ is monotone–that is, has a strictly positive kernel–the left member of (38) is strictly positive if the operand is so. Therefore, by induction on n, $F_{n+1} < F_n \ \forall n$.

The sequence $\{F_n(t)\}$ is strictly decreasing and bounded below by zero. Thus it has a limit $L(t)$, and this defines a function L on $[0, \infty)$ such that $F_n \to L$ as $n \to \infty$. By Lemma 4, L is stationary under $\mathfrak{M}_\lambda$. Thus L must be either 0, one of the F_λ, or F_λ^T.

Let H be any solution of (12), so that $H = \mathfrak{M}_\lambda H^m$. If $F_n > H$ then $F_n^m > H^m$, so that $F_{n+1} > H$ because, as in (38), $F_{n+1} - H = \mathfrak{M}_\lambda(F_n^m - H^m)$. Every solution of (12), being unlimited, is less than 1, so that $H < F_1$. By induction on n, $H < F_n \ \forall n$. That is, H is a pointwise lower bound for $\{F_n\}$. As we found on p. 1388, the largest solution of (12) is F_λ^T, so that $F_\lambda^T(t)$ is the greatest of the lower bounds represented by $H(t)$. Therefore $L = F_\lambda^T$. □

We can write Theorem 2 as the statement that $F_n[1, \lambda] \downarrow F_\lambda^T$ when $\lambda \leqq \Lambda$. (Actually the convergence is uniform in this case, but there is no need to prove this.) This result is in essence the answer to our problem, although we must now generalize it considerably in order to remove the strong restriction that $F_1 = 1$.

Before going on to strengthen Theorem 2, we observe that receiver delays increase indefinitely if the system continues to operate with more-than-critical load. This fact is stated accurately in

Theorem 3: Choose a fixed $\lambda > \Lambda(m)$. Then the sequence $\{F_n[1, \lambda]\}$ approaches zero.

Proof: The proof of Theorem 2 applies here with one exception. Again $\{F_n\}$ has a limit L which is stationary under $\mathfrak{M}_\lambda$, but with $\lambda > \Lambda$ the only solution of (12) is identically zero. □

In this case the convergence is not uniform; the probability masses associated with successive members of $\{F_n\}$ are located farther and farther to the right. We may conclude our discussion of the supercritical case with

Corollary 1: Let F_1 be any d.f. and choose a fixed $\lambda > \Lambda(m)$. Then the sequence $\{F_n[F_1, \lambda]\}$ approaches zero.

Proof: Since F_1 is a d.f., $F_1 \leqq 1$. Thus $F_1^m \leqq 1^m$, and so $F_2[F_1, \lambda] - F_2[1, \lambda] = \mathfrak{M}_\lambda(F_1^m - 1^m) \leqq 0$. Likewise $F_n[F_1, \lambda] \leqq F_n[1, \lambda] \ \forall n$ by induction. But $F_n[F_1, \lambda] \geqq 0$, and $F_n[1, \lambda] \to 0$ as $n \to \infty$ by Theorem 3. Therefore $F_n[F_1, \lambda] \to 0$ as $n \to \infty$. □

Thus, no matter what the delay distribution encountered by the first message, delays increase without bound when λ exceeds Λ.

Our first step in generalizing Theorem 2 is to consider the case in which the system, instead of starting empty, is first allowed to reach statistical equilibrium with a receiver arrival-rate $\mu \leqq \Lambda$. In effect, the system starts operation at time $-\infty$ and comes to equilibrium before the arrival (in receiver queues) of the message we label number one. At that instant we change the receiver arrival-rate to a new value $\lambda \leqq \Lambda$ and thenceforth keep it fixed. This mathematical model is a very idealized one, since it is not clear how the receiver interarrival-times can suddenly acquire the density $\lambda \exp(-\lambda t)$ at a specified instant. With k and m fixed, the arrival rate can change from μ to λ only through a change in the rate α of transmitter arrivals; and the effect of a change in α will be felt gradually as the number of occupied transmitters increases to its new equilibrium-value. However, since eqs. (7) and (36) form our only tool for studying the evolution of receiver delays, the best we can do in the present case is to take $F_1 = F_\mu^T$ (by Theorem 2) and to generate $\{F_n[F_\mu^T, \lambda]\}$ by applying $\mathfrak{M}_\lambda$. We begin with

Lemma 5: If 0 < $\lambda < \mu < \Lambda(m)$, then 0 > $\gamma_2(\lambda) > \gamma_2(\mu) > \gamma_1(\mu) > \gamma_1(\lambda) > -1$. If $\mu = \Lambda$, then $\gamma_2(\mu) = \gamma_1(\mu) \equiv \gamma(\mu) = (\Lambda - 1)/2$. If $\lambda = 0$, then $\gamma_2(\lambda) = 0$ and $\gamma_1(\lambda) = -1$.

Proof: The result for $\mu = \Lambda$ was stated on p. 1382. The result for $\lambda = 0$ is obvious from (25). From (25),

$$2\frac{d\gamma_1}{d\lambda} = 1 + \frac{2m - (1 + \lambda)}{[(1 + \lambda)^2 - 4\lambda m]^{\frac{1}{2}}}$$

which is clearly positive because $m > 1$ and $\lambda \leqq \Lambda < 1$. Thus $\mu > \lambda$ implies that $\gamma_1(\mu) > \gamma_1(\lambda) > -1$. By definition $\gamma_2(\mu) > \gamma_1(\mu)$ when $\mu < \Lambda$. To show that $0 > \gamma_2(\lambda) > \gamma_2(\mu)$, we need only prove that

$$2\frac{d\gamma_2}{d\lambda} = 1 - \frac{2m - (1 + \lambda)}{[(1 + \lambda)^2 - 4\lambda m]^{\frac{1}{2}}} < 0,$$

and this follows after trivial manipulation from the fact that $m > 1$. □

The next lemma expresses another kind of "monotonicity" property of $\mathfrak{M}_\lambda$, this time with respect to the parameter λ.

Lemma 6: Let H_μ be any nonzero d.f. stationary under $\mathfrak{M}_\mu$. If $\lambda < \mu$, then $\mathfrak{M}_\lambda H_\mu > H_\mu$; and if $\lambda > \mu$, then $\mathfrak{M}_\lambda H_\mu < H_\mu$.

Proof: We express the action of $\mathfrak{M}_\lambda$ as in eqs. (35), with k_λ given by (14). First assume that $\lambda < \mu$, and define τ as the value of u at which

k_λ and k_μ intersect; see Fig. 5. That is, $k_\lambda(t, \tau) = k_\mu(t, \tau)$, so that τ is a function of t. [Explicitly, $\tau = t + (\mu - \lambda)^{-1} \cdot \ln([1 + \lambda^{-1}]/[1 + \mu^{-1}])$.] It is easy to see that, as indicated in Fig. 5, $k_\lambda(t, u) < k_\mu(t, u)$ when $u < \tau$ and $k_\lambda(t, u) > k_\mu(t, u)$ when $u > \tau$.

For any fixed $t \geqq 0$,

$$\begin{aligned} \mathfrak{M}_\lambda H_\mu(t) - H_\mu(t) &= \mathfrak{M}_\lambda H_\mu(t) - \mathfrak{M}_\mu H_\mu(t) \\ &= [(\mathfrak{M}_\lambda - \mathfrak{M}_\mu)H_\mu^m](t) \\ &= \int_0^\infty [k_\lambda(t, u) - k_\mu(t, u)]H_\mu^m(u)\, du \\ &= \int_\tau^\infty [k_\lambda(t, u) - k_\mu(t, u)]H_\mu^m(u)\, du \\ &\quad - \int_0^\tau [k_\mu(t, u) - k_\lambda(t, u)]H_\mu^m(u)\, du. \end{aligned}$$

We know that in each of the integrals the bracketed quantity is positive. Also H_μ is nondecreasing; thus $H_\mu(\tau) \leqq H_\mu(u)$ for $u > \tau$ and $H_\mu(\tau) \geqq H_\mu(u)$ for $u < \tau$. Therefore

$$\begin{aligned} \mathfrak{M}_\lambda H_\mu(t) - H_\mu(t) &\geqq H_\mu^m(\tau)\left\{\int_\tau^\infty [k_\lambda(t, u) - k_\mu(t, u)]\, du \right. \\ &\qquad \left. - \int_0^\tau [k_\mu(t, u) - k_\lambda(t, u)]\, du\right\} \\ &= H_\mu^m(\tau) \int_0^\infty [k_\lambda(t, u) - k_\mu(t, u)]\, du \\ &= H_\mu^m(\tau)\{\mathfrak{M}_\lambda(1) - \mathfrak{M}_\mu(1)\} \\ &= H_\mu^m(\tau)e^{-t}\left\{\frac{\mu}{1+\mu} - \frac{\lambda}{1+\lambda}\right\}, \end{aligned}$$

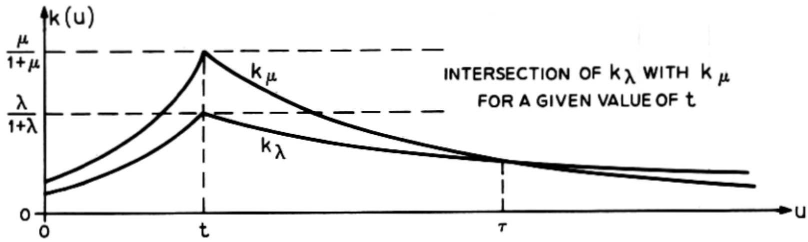

Fig. 5—The kernels k_λ and k_μ cross at τ.

where the last equality comes from (13) as noted in the proof of Theorem 2. The explicit formula for τ shows that $\tau > 0 \ \forall t$, so that $H_\mu(\tau) > 0$ because H_μ is nonzero. [By the equilibrium form of (8) on p. 1375, $H_\mu(0) > 0$.] Since $\lambda < \mu$, each factor of the last expression displayed above is positive, and thus $\mathfrak{N}_\lambda H_\mu(t) > H_\mu(t) \ \forall t$. This proves the first conclusion of the lemma.

When $\lambda > \mu$, the proof goes through as before except that the appropriate inequalities are reversed. The modified proof shows that $\mathfrak{N}_\lambda H_\mu(t) - H_\mu(t)$ is less than or equal to the final expression in the calculation above, and that expression is negative in this case. □

Lemma 7: If $\lambda < \mu \leqq \Lambda(m)$, *then* $F_\lambda^T > F_\mu^T$.

Proof: Because F_μ^T is unlimited, $1 > F_\mu^T$. Applying $\mathfrak{N}_\lambda$, we find that $F_2[1, \lambda] \equiv \mathfrak{N}_\lambda 1 > \mathfrak{N}_\lambda F_\mu^T > F_\mu^T$, where the first inequality comes from the argument used with (38) and the second from Lemma 6. Repeated application of $\mathfrak{N}_\lambda$ shows that $F_n[1, \lambda] > \mathfrak{N}_\lambda F_\mu^T \ \forall n$, so that, using Theorem 2,

$$F_\lambda^T \equiv \lim_{n\to\infty} F_n[1, \lambda] \geqq \mathfrak{N}_\lambda F_\mu^T > F_\mu^T . \ \square$$

Lemma 8: Let F be any solution of (12), and let the corresponding phase-plane trajectory $p(F)$ *reach* $(F, p) = (1, 0)$ *with slope* $\gamma < 0$. *Then for any* $\epsilon > 0$, *there exist constants* c_3, c_4, *and T such that, whenever* $t > T$,

$$c_3 e^{(\gamma-\epsilon)t} < 1 - F(t) < c_4 e^{(\gamma+\epsilon)t}. \tag{39}$$

Proof: From (19), $dt = dF/p(F)$, so that we can write

$$t(F) = \int_{F_0}^{F} \frac{du}{p(u)}. \tag{40}$$

Given $\epsilon > 0$, there is an abscissa F_ϵ such that the integral curve $p(F)$ lies in the wedge $-(\gamma + \epsilon)(1 - F) < p < -(\gamma - \epsilon)(1 - F)$ when $F_\epsilon < F \leqq 1$. Choose T so large that $F(T) > F_\epsilon$. For $t > T$, (40) becomes

$$t(F) = \int_{F_0}^{F(T)} \frac{du}{p(u)} + \int_{F(T)}^{F} \frac{du}{p(u)},$$

and the first of these integrals equals T by definition. When $F > F(T)$, $p(F)$ satisfies the inequalities defining the wedge just mentioned, and so

$$T - \int_{F(T)}^{F} \frac{du}{(\gamma - \epsilon)(1 - u)} < t(F) < T - \int_{F(T)}^{F} \frac{du}{(\gamma + \epsilon)(1 - u)}.$$

Thus

$$T + \frac{1}{\gamma - \epsilon} \ln \frac{1 - F(t)}{1 - F(T)} < t < T + \frac{1}{\gamma + \epsilon} \ln \frac{1 - F(t)}{1 - F(T)},$$

and the conclusion (39) follows after rearrangement. The constants are

$$c_3 = [1 - F(T)]e^{-(\gamma - \epsilon)T}, \quad c_4 = [1 - F(T)]e^{-(\gamma + \epsilon)T}. \quad \square$$

[The facts that $1 - F(t)$ behaves nearly exponentially for large t, that the decay constant γ must take one of the forms (25), and that no admissible solution exists when $\lambda > \Lambda(m)$, can all be inferred non-rigorously by considering linearized versions of (12a), (18a), or (20a) for very small values of $1 - F$.]

Lemma 9: When $\lambda < \mu$, no member of the open family $\{F_\lambda\}$ lies above F^T_μ.

Proof: For any given F_λ and $\epsilon > 0$ and all sufficiently large t, we know from Lemma 8 that there exist constants c_λ and c_μ such that

$$F_\lambda(t) < 1 - c_\lambda e^{[\gamma_2(\lambda) - \epsilon]t} \quad \text{and} \quad F^T_\mu(t) > 1 - c_\mu e^{[\gamma_1(\mu) + \epsilon]t}.$$

By Lemma 5, $\gamma_2(\lambda) > \gamma_1(\mu)$. So long as the ϵ that appears above is small enough to make $\gamma_2(\lambda) - \epsilon > \gamma_1(\mu) + \epsilon$, we can find τ such that $F^T_\mu(t) > F_\lambda(t)$ when $t > \tau$. $\square$

Theorem 4: Let $\lambda \leqq \Lambda(m)$ and $\mu \leqq \Lambda(m)$. Then $F_n[F^T_\mu, \lambda] \rightarrow F^T_\lambda$ as $n \rightarrow \infty$.

Proof: If $\lambda = \mu$, the result is obvious because F^T_μ is stationary under $\mathfrak{N}_\mu$.

If $\lambda > \mu$, then $F_2[F^T_\mu, \lambda] \equiv \mathfrak{N}_\lambda F^T_\mu < F^T_\mu$ by Lemma 6. Therefore, by the argument based on monotonicity and induction that was used in proving Theorem 2, the sequence $\{F_n[F^T_\mu, \lambda]\}$ is strictly decreasing. Thus it has a limit $L \geqq 0$ which by Lemma 4 is stationary under $\mathfrak{N}_\lambda$. Exchanging λ with μ in Lemma 7, we see that $F^T_\lambda < F^T_\mu$. Continuing the argument of Theorem 2, we see that F^T_λ is a lower bound for the sequence $\{F_n[F^T_\mu, \lambda]\}$, and greater than any of the other lower bounds F_λ. Thus $F_n[F^T_\mu, \lambda] \downarrow F^T_\lambda$.

If $\lambda < \mu$, a similar argument shows that F^T_λ is an upper bound for the strictly increasing sequence $\{F_n[F^T_\mu, \lambda]\}$, which again has a limit stationary under $\mathfrak{N}_\lambda$. By Lemma 9, no other function F_λ that is stationary under $\mathfrak{N}_\lambda$ can be an upper bound for this sequence. Therefore $F_n[F^T_\mu, \lambda] \uparrow F^T_\lambda$. $\square$

It is natural to consider the communication system as initially idle and thenceforth, after the arrival of the first message, subjected to a load of constant intensity. Theorem 4 generalizes this situation by considering changes from a previously attained equilibrium state. But the most realistic and therefore useful statement about limiting distributions would involve initialization of some nearly arbitrary kind, so that any queue-length distribution might be encountered by message number one. We now study a model of this kind in two lemmata and a theorem. In this model the system begins empty, receives messages at a sequence of different arrival-rates which may be supercritical, and is then subjected to a single subcritical load indefinitely. A real system operating with time-varying load would experience a sequence of receiver interarrival-times with different and presumably non-exponential distributions. We consider an idealized counterpart in which these distributions are indeed exponential.

Lemma 10: Let the sequence $\{F_n[1, \{\lambda_i\}]\}$ be generated by applying $\mathfrak{N}_{\lambda_i}$ k_i times, $i = 1, \cdots, j$, so that $n = 1 + \sum_{i=1}^{j} k_i$, with each $\lambda_i > 0$. Then $F_n[1, \{\lambda_i\}](t) \cong 1 - c_n t^{n-2} \exp(-t)$.

Proof: We suppose that the exact expression for F_n is a series beginning with the two terms shown and continuing with exp $(-t)$ times lower powers of t, exp $(-2t)$ times higher-degree polynomials in t, and so on. We know that $F_2[1, \lambda_1] = 1 - [\lambda_1/(1 + \lambda_1)] \exp(-t)$, which has the assumed form. Also

$$F_n[1, \{\lambda_i\}] = \mathfrak{N}_{\lambda_j}^{k_j} \mathfrak{N}_{\lambda_{j-1}}^{k_{j-1}} \cdots \mathfrak{N}_{\lambda_1}^{k_1}(1) = \mathfrak{N}_{\lambda_j} F_{n-1}[1, \{\lambda_i\}].$$

Assume $F_{n-1}[1, \{\lambda_i\}]$ has the stated form, substitute into (7), expand F_{n-1}^m using the binomial theorem, and integrate explicitly. This very tedious procedure, which, being straightforward, is not recorded here, yields the conclusion of the lemma by induction. The constant c_n is a function of m and of the λ_i. □

Corollary 2: The conclusion of Lemma 10 holds if, in generating $\{F_n\}$, the receiver arrival-rate is set equal to zero during a finite number of finite intervals.

Proof: If the load is zero for an interval of length t_0 between the arrivals of messages $n - 1$ and n, then F_n is simply shifted t_0 units to the left; that is, the probability that would have been $F_n[1, \{\lambda_i\}](t)$ becomes $F_n[1, \{\lambda_i\}](t - t_0)$. This expression, with F_n calculated as in Lemma 10, defines a new function of t having the same form as $F_n[1, \{\lambda_i\}]$ but with different constants. This process does not affect the proof of Lemma 10 and can be repeated finitely many times. □

Note that the loads $\{\lambda_i\}$ in Lemma 10 need not be less than or equal to $\Lambda(m)$. In describing transient loads, we measure the length of time during which λ_i is the receiver arrival-rate by specifying k_i, the number of messages which follow their predecessors by intervals with the density $\lambda_i \exp(-\lambda_i t)$. Since this procedure cannot account for a period with zero arrival-rate, we must assume $\lambda_i > 0$ in Lemma 10 and then show separately in Corollary 2 that the form of the sequence $\{F_n[1, \{\lambda_i\}]\}$ is not affected by the existence of periods during which no messages can arrive. The effect of shutting off transmitter arrivals in an actual system would be approximated by a sequence of λ_i gradually decreasing to zero as the transmitter queues empty out.

Lemma 11: Let F be any element $F_n[1, \{\lambda_i\}]$ of the sequence defined in Lemma 10. Then there exists a member F_Λ, of the open family of d.f.s stationary under $\mathfrak{M}_\Lambda$, such that $F_\Lambda < F$.

Proof: We know from Lemma 5 that $\gamma(\Lambda) > -1$; choose $\epsilon < [1 + \gamma(\Lambda)]/2$, so that $\gamma(\Lambda) - \epsilon > -1 + \epsilon$. According to (39) we can find τ_1 and c such that

$$F_\Lambda^T(t) < 1 - ce^{[\gamma(\Lambda)-\epsilon]t} \quad \text{for} \quad t > \tau_1 .$$

Also, by Lemma 10, we can choose τ_2 and c_n such that

$$F(t) > 1 - c_n t^{n-2} e^{-(1-\epsilon)t} \quad \text{for} \quad t > \tau_2 ,$$

since for large enough t, the error term in Lemma 10 can be made small enough to be bounded by the effect of the factor $\exp(\epsilon t)$. (Here n is the index of F in $\{F_n\}$.) There exists τ_3 such that $c_n t^{n-2}/c < \exp([1 + \gamma(\Lambda) - 2\epsilon]t)$ when $t > \tau_3$. Let $\tau = \max(\tau_1, \tau_2, \tau_3)$. Then $F_\Lambda^T(t) < F(t)$ for $t > \tau$. Since $F_\Lambda < F_\Lambda^T$ for every member of the family $\{F_\Lambda\}$, we have $F_\Lambda(t) < F(t)$ when $t > \tau$ for each F_Λ–that is, uniformly in the parameter $F_\Lambda(0)$ that we use to specify an element of $\{F_\Lambda\}$.

By Theorem 1, every element of the sequence $\{F_n\}$ is proper, and so by (8) $F(0) > 0$. Choose $F_\Lambda(0) < F(0) \exp(-\Lambda\tau)$. From the equilibrium form of (9) we have $F_\Lambda(\tau) < F_\Lambda(0) \exp(\Lambda\tau) < F(0)$. Since F_Λ and F are nondecreasing, $F_\Lambda(t) < F(t)$ when $t \leqq \tau$. We have now proved this inequality for all t, so that $F_\Lambda < F$. □

Theorem 5: Let F be any element $F_n[1, \{\lambda_i\}]$ of the sequence defined in Lemma 10. Choose $\lambda < \Lambda(m)$. Then $F_n[F, \lambda] \to F_\lambda^T$ as $n \to \infty$.

Proof: Given F, choose F_Λ by Lemma 11 so that $F_\Lambda < F$. By Lemma 5, $\gamma_2(\lambda) > \gamma(\Lambda)$. Thus Lemma 9 remains valid with F_μ^T replaced by any member of $\{F_\Lambda\}$. Therefore Theorem 4 is valid with F_μ^T replaced by F_Λ, so that $F_n[F_\Lambda, \lambda] \uparrow F_\lambda^T$. By Theorem 2, $F_n[1, \lambda] \downarrow F_\lambda^T$. By Theorem 1, F is unlimited, and we can write $F_\Lambda < F < 1$. By the monotonicity argument of Theorem 2, $F_n[F_\Lambda, \lambda] < F_n[F, \lambda] < F_n[1, \lambda]$ for all n. Therefore $\{F_n[F, \lambda]\}$ is bounded term by term on both sides by sequences which approach F_λ^T, and so $F_n[F, \lambda] \to F_\lambda^T$ as $n \to \infty$. □

This theorem does not cover the case of a system operated *at* its critical load after being subjected to an arbitrary transient of finite duration. Part of the gap is filled by

Corollary 3: Let F be chosen as in Theorem 5 but with $F \geqq F_\Lambda^T$, *for example because max* $\{\lambda_i\} \leqq \Lambda$. *Then* $F_n[F, \Lambda] \to F_\Lambda^T$ *as* $n \to \infty$.

Proof: $F_\Lambda^T \leqq F < 1$; $F_n[1, \Lambda] \downarrow F_\Lambda^T$ by Theorem 2; and $F_n[F_\Lambda^T, \Lambda] = F_\Lambda^T \ \forall n$. Thus $\{F_n[F, \Lambda]\}$ is squeezed between two sequences with the same limit. □

This kind of argument does not work when F is not greater than or equal to F_Λ^T. We can bound F below by a member of the family $\{F_\Lambda\}$, but F_Λ is stationary under $\mathfrak{N}_\Lambda$; thus, if $\{F_n[F, \Lambda]\}$ or a subsequence of it has a limit L, we know only that $\forall t \ L(t)$ lies in the interval $[F_\Lambda(t), F_\Lambda^T(t)]$. We could choose $\lambda < \Lambda$ and bound F below by a particular F_λ, but this would not help because $\mathfrak{N}_\Lambda F_\lambda < F_\lambda$ by Lemma 6: $\{F_n[F, \Lambda]\}$ would be bounded below by a decreasing sequence. I have not been able to determine the behavior of the sequences in question more precisely than is stated in

Corollary 4: Let F be chosen as in Theorem 5 but with F not bounded below by F_Λ^T. *Then there exists an* $F_\Lambda \in \{F_\Lambda\}$: $F_n[F, \Lambda] > F_\Lambda \ \forall n$; *and if* $\{F_n[F, \Lambda]\}$ *has a limit L, then L is either* F_Λ^T *or a member of* $\{F_\Lambda\}$.

Proof: By Lemma 11 we can find $F_\Lambda < F$, and by the monotonicity of $\mathfrak{N}_\Lambda$, this inequality is preserved throughout the sequence $\{F_n[F, \Lambda]\}$. By Lemma 4, L is stationary under $\mathfrak{N}_\Lambda$; and our first conclusion shows that L is bounded away from zero. □

The limitations of this result are not surprising. If the system is operated *at* its critical load after being temporarily overloaded, we see that delays still do not increase indefinitely; but on the other hand we have found no assurance that there is a limiting distribution or that, if there is, it agrees with the one (F_Λ^T) that would have resulted if the period of critical load had begun with the system idle.

We defer discussion of all these results to later sections, merely observing that the argument given on pp. 1388–9 makes Theorems 2, 4, and 5 quite remarkable. Another interesting point is raised by the strong dependence of these proofs on the *asymptotic* behavior of the functions involved–in particular, on the rate at which $F_1(t) \to 1$ for large t, where F_1 is the initial distribution of a sequence $\{F_n[F_1, \lambda]\}$. It may seem that our results are too sensitive to such decay rates–that this feature somehow represents an "instability" of the mathematical model. But in fact the essentials are the general properties recorded in Theorem 1, and the decay behavior of F_n which Lemma 10 shows is dominated by $t^{n-2} \exp(-t)$. All these properties are in turn controlled by the density $\exp(-t)$ of the transmission times, and the expression $t^{n-2} \exp(-t)$ in particular is related to the fact that the sum of the $n - 1$ preceding transmission times *or their tails* has an Erlang distribution of order $n - 1$. The most interesting implication of the dependence of our arguments on asymptotic behavior is that our results are accessible only to analytic techniques: A computer could not be successfully used to study experimentally the properties of sequences generated by eq. (7).

The situation for $m = 1$ puts the argument of this section in perspective. All solutions of (12) and (18) have the form $F(t) = C[1 - \lambda \exp(-(1 - \lambda)t)]$ in this case, with $0 < C \leqq 1$. Only the one of these with $C = 1$ is proper, and it is Erlang's well-known delay-distribution for the single-server queue.

V. DISTRIBUTION OF RECEIVER DELAYS

We have found that the meaningful solutions of the stationarity equations, (12) and (18), are the limiting distributions F_λ^T. This section is devoted to four questions: How can these distributions be calculated explicitly; how do we proceed when the number of addresses per message is a random variable of which m is the mean; what are the properties of the delay distributions; and how closely do these results represent the behavior of systems with finite R?

5.1 *Computation of Delay Distributions and Their Moments*

In answering the first of these questions we take λ and m to be fixed and interpret the symbol F to mean F_λ^T; likewise the symbol γ_1 refers to the quantity defined in (25). The first step is to carry out the first integration, based on

$$p' = [\{\lambda(F - F^m)\}/p] - (1 - \lambda), \tag{20a}$$

which yields the phase-plane trajectory $p(F)$. (See Fig. 4.) This can be done numerically using any standard integration-formula of the predictor-corrector type, and is best done as follows: Beginning at the corner (1, 0), integrate the vector field to the left, starting at an angle with the F-axis whose tangent is γ_1. Follow the resulting path until it intersects the edge $p = \lambda F$ at the abscissa $\hat{F}_0$, which is an estimate of the true initial value $F(0)$ of the distribution sought. Since the vector field diverges in the leftward direction, this integration is numerically quite unstable, so that $\hat{F}_0$ need not be an accurate estimate of $F(0)$. Explore the neighborhood of $\hat{F}_0$ for the true value of $F(0)$ by selecting a set of abscissae (spaced about 10^{-5} units apart) and calculating, for each one, the integral curve that passes through it. These calculations are performed by integrating to the right; and since the vector field converges to the right, the resulting curves are quite accurate. For each such curve, find the ordinate at which it intersects the line $F = 1$. These ordinates can be plotted against the abscissae at which the integral curves start on the edge $p = \lambda F$. Beginning at some point well to the right of $\hat{F}_0$, the graph so constructed has an ordinate of considerable positive magnitude. As the abscissa decreases, approaching the true $F(0)$ from above, the ordinate must decrease to zero, remaining there for all lower values of the abscissa; for trajectories beginning to the left of F_0^T must all end at (1, 0). This ideal pattern is perturbed by noise arising from roundoff and from the numerical integration itself, but it is not difficult to find $F(0)$ from this graph of p-intercepts to an accuracy ranging from about 10^{-4} to about 10^{-6} units, so long as λ is not very close to Λ. The function $p(F)$ is then found by integrating the vector field to the right from $F(0)$. [Trajectories lying below F_λ^T cannot be found precisely; they have very large curvatures near (1, 0) because of the singularity there.]

This technique must be modified when $\lambda = \Lambda(m)$ (and in fact when $\lambda/\Lambda \gtrsim 0.9$), for in this case the graph just described is erratic and appears to have quite a high-order contact with its abscissa, the F_0-axis. Less than two decimal-places of precision can be obtained in this way. Barbara R. LaCava suggested looking for the trajectory which has the smallest number of corrector cycles per predictor step in integrating leftward; and this yields an order-of-magnitude improvement. In order to find $F(0)$ to five-place accuracy, we had to obtain an accurate value of $p(F)$ analytically for $F \cong 0.9998$ and to use numerical techniques only for smaller values of F. The analytical values came from a pair of parametric power-series about (1, 0) for the coordinates of the desired trajectory. Such series, describing trajectories in the neighborhood of an improper node, can be found from the method of Picard,[7] of which the existence and relevance were pointed out to me by

A. Kuczura. (This method is also mentioned by Kaplan,[8] who does not cover the case of $\lambda = \Lambda$ in which $\gamma_1 = \gamma_2$.)

The curve $p(F)$ can be parametrized by numerical evaluation of the integral

$$t(F) = \int_{F_0}^{F} \frac{du}{p(u)}. \tag{40}$$

This process yields the inverse $t(F)$ of the desired distribution function. As $(F, p) \to (1, 0)$, the numerical integration becomes inaccurate and must be replaced by analytical approximation. The values $F(t)$ for large t can be estimated by the exponential obtained from (40) by assuming that $p(F)$ represents a straight line to the right of some point (F, p). This can be taken, for example, as the straight line of slope γ_1 that passes through (1, 0) or as the chord from (F, p) to (1, 0).

These procedures* yield a table of triples (F, p, t) from which the distribution F and the density p can be plotted, and the sojourn-time distribution G and its transform computed. The moments M_i of the receiver delay about zero can also be found from $p(F)$, as follows: By definition,

$$M_i(F) \equiv E(\omega^i) \equiv \int_0^\infty t^i \, dF(t) = i \int_0^\infty t^{i-1}[1 - F(t)] \, dt, \tag{41}$$

where E is the expectation operator. The integration by parts that leads to the final expression above is valid whenever $1 - F(t)$ goes to zero faster than t^{-i} as $t \to \infty$. By Lemma 8 this is true $\forall i$ if $\gamma_1 < 0$, as is always the case for $0 < \lambda \leqq \Lambda(m)$ according to Lemma 5. (Indeed, we see that F has finite moments of all orders, as do the members of the other family $\{F_\lambda\}$ of stationary distributions.) We can rewrite (41) as an integral with respect to F, using $dt = dF/p(F)$ from (19):

$$M_i(F) = i \int_{F_0}^{1} [t(F)]^{i-1}\{(1 - F)/p(F)\} \, dF. \tag{42}$$

Because these integrals are hard to evaluate precisely, we divide them into two parts as shown next for the case of M_2. From (42),

$$M_2(F) = 2 \int_{F_0}^{F(T)} t(F) \frac{1 - F}{p(F)} \, dF + 2 \int_{F(T)}^{1} t(F) \frac{1 - F}{p(F)} \, dF.$$

Following the proof of Lemma 8, for sufficiently large T we can write the approximation

* A simpler and more accurate method can be used when m is an integer.

$$t(F) \cong T + \frac{1}{\gamma_1} \ln \frac{1 - F}{1 - F(T)}.$$

Then the last integral above becomes

$$-2 \int_{F(T)}^{1} \left[T + \frac{1}{\gamma_1} \ln \frac{1 - F}{1 - F(T)} \right] \frac{dF}{\gamma_1},$$

where we have replaced $(1 - F)/p(F)$ by $-1/\gamma_1$ because the approximation $p(F) \cong -\gamma_1(1 - F)$ improves as $p(F) \to 0$. This integral can be written as

$$-\frac{2}{\gamma_1} [1 - F(T)] \left[T + \frac{1}{\gamma_1} \int_0^1 \ln u \, du \right],$$

and so M_2 becomes

$$M_2(F) \cong 2 \int_{F_0}^{F(T)} t(F) \frac{1 - F}{p(F)} \, dF - \frac{2}{\gamma_1} [1 - F(T)][T - (1/\gamma_1)]. \qquad (43)$$

The integral term can be calculated from the previous evaluation of $F(t)$. As with F, we get a slightly different estimate of M_2 by using the slope $p(F(T))/[1 - F(T)]$ of the chord, in place of γ_1.

Also, we have not yet specified the value T of the time-parameter at which the integral defining $M_2(F)$ is "broken in two." Each such choice leads to a particular estimate of M_2. If T is too small, $p(F)$ is not accurately approximated by a straight line to the right of $(F(T), p(F(T)))$; but as T increases, accurate evaluation of the integral in (43) becomes more difficult. Thus it is useful to evaluate (43) for many values of T. As T increases, at first the calculated values of M_2 smoothly approach a limit which can be estimated graphically; subsequently these values start to behave erratically as the numerical integration loses precision. (A similar method applies in the simpler case of M_1.)

5.2 *Variable Number of Addresses per Message*

The number m, which we first took to be an integer, enters the problem only through the function F_n^m in (3), by way of the definition

$$\omega + \xi = \max_{i \leq m} \omega^{(i)} \qquad (2)$$

and the property that the random variables $\omega^{(i)}$ are independent. But suppose instead that the number of addresses of a message takes the value j with probability ν_j, independently of the numbers of addresses of all other messages; and that

$$m = E(j) \equiv \sum_{j=1}^{\infty} j\nu_j \tag{44}$$

is the average multiplicity. Then with probability ν_j we have $\omega + \xi = \max_{i \leq j} \omega^{(i)}$, so that the distribution of $\omega + \xi$ is F^j with probability ν_j. Thus the true distribution of $\omega_n + \xi_n$ that appears in (3) should be

$$\sum_{j=1}^{\infty} \nu_j F_n^j ,$$

which also appears in the recurrences (7) and (37) and leads to the equilibrium equation

$$F = \mathfrak{M}_\lambda(\Sigma_1^\infty \nu_j F^j). \tag{45}$$

[The operator $\mathfrak{M}_\lambda$ is defined in (35b).]

This equation fits nicely into the framework of our earlier results. We now think of the distribution of $\omega + \xi$ as being represented by a function $\Delta(F)$, defined by

$$\Delta(F) = \Sigma_1^\infty \nu_j F^j \tag{46}$$

rather than by the distribution F^m as in Section 2.3. Then the phase-plane differential equation (20a) becomes

$$p'(F) = \frac{\lambda}{p}[F - \Delta(F)] - (1 - \lambda), \tag{47}$$

which leads to an analogue of (22) with m replaced by $\Delta'(1)$ in the linear term and each binomial coefficient $\binom{m}{j}$ replaced by a multiple of the jth derivative $\Delta^{(j)}(1)$ (which is assumed to exist). The analysis of Cases 1 through 5 proceeds as before. In particular, eqs. (25) and (26a) [for γ_1 and $\Lambda(m)$] are valid with m replaced by $\Delta'(1)$. But here, by (46) and (44), $\Delta'(1) = m$; and so the critical load Λ and the limiting slope γ_1 are meaningful and correct even when m is merely the *average* number of addresses per message! It is this result that justifies treating m as a real number exceeding 1 in the analysis of Sections III and IV, in contrast to the original appearance of m as in integer in eqs. (2) and (3).

We may obtain explicit numerical results for a distribution $\{\nu_j\}$ by means of the procedure of Section 5.1, using (47) to find the phase-plane trajectory. No new difficulty is encountered, because we know γ_1 exactly. It is convenient to have a simple expression for $\Delta(F)$, as is possible when $\{\nu_j\}$ has a form allowing explicit summation of the series in (46). For example, when $\{\nu_j\}$ is geometric with parameter q, so that $m = (1 - q)^{-1}$, we find that $\Delta(F) = F/[m(1 - qF)]$.

Unless $\nu_j = \delta_{jm}$, we know that $\Delta(F) > F^m$ when $E(j) = m$ in (46). It follows from (20a) and (47) that $F^T_{\lambda,\Delta} > F^T_{\lambda,m}$, where $F^T_{\lambda,\Delta}$ is the (uppermost) solution of (45) and $F^T_{\lambda,m}$ that of (12) and (20). For our purposes, in view of the invariance of Λ and γ_1 under m-preserving changes in $\{\nu_j\}$, it is sufficient to take $F^T_{\lambda,m}$ as an approximation to $F^T_{\lambda,\Delta}$; the latter could be found numerically if needed for engineering purposes. We could also construct a quantitative theory for approximating $F^T_{\lambda,\Delta}$ by some other $F^T_{\lambda,M}$ with properly chosen M, proceeding by way of (47), (40), and this analogue of (38) (with superscript T suppressed):

$$F_{\lambda,m} - F_{\lambda,\Delta} = \mathfrak{M}_\lambda(F^m_{\lambda,m} - \Delta(F_{\lambda,\Delta})).$$

Such an investigation does not seem worthwhile; we merely note that certainly $M < m$.

5.3 *Numerical Results*

We begin by examining the function

$$\Lambda(m) = 2m - 1 - 2(m^2 - m)^{\frac{1}{2}} \tag{26a}$$

that specifies the maximum rate of receiver arrivals that allows of an equilibrium delay-distribution. This function [eq. (26a)] is plotted in Fig. 6. Its most striking property is its rapid decrease as m increases from 1: Indeed, the slope $d\Lambda/dm$ is $-\infty$ at $m = 1+$. The critical load Λ is down to 0.5 at $m = 1.125$ (corresponding, for example, to one-eighth of the messages having two addresses and all the rest, one); other values are $\Lambda(2) \cong 0.172$, $\Lambda(3) \cong 0.101$, and $\Lambda(10) \cong 0.0263$. Since m ranges from 2 to 3 in a number of practical situations, we see how severely the camp-on discipline limits the possible efficiency of a very large multiple-address system. As discussed in Section VII,

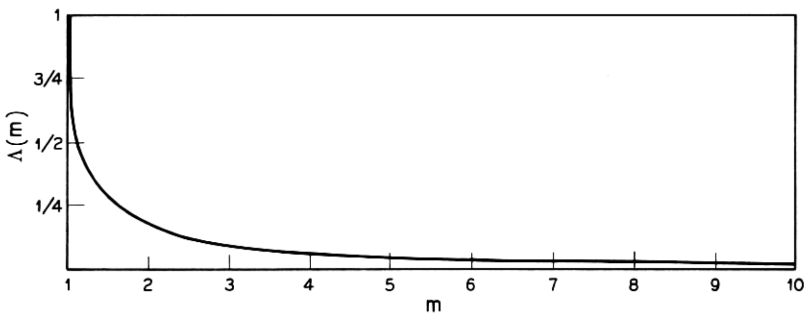

Fig. 6—The critical load for m addresses per message.

this fact was known qualitatively; but explicit knowledge of the function $\Lambda(m)$ is new. It suggests, as mentioned in Section I, that in practice a good system is likely to employ a mixed discipline.

Eight typical functions $F^T_{\lambda,m}$ are shown in Figs. 7–9, as computed–with considerable perseverance, necessitated by the singularity at $(F, p) = (1, 0)$–by LaCava. Figure 7 shows the receiver-delay densities for three values of λ when $m = 2$ addresses per message. The logarithm of the probability density $\dot{F}^T_\lambda$ is plotted against time in units of the mean length of a message, as in Figs. 8 and 9. [The probability F_0 of no delay, and the relative traffic intensity $\lambda/\Lambda(m)$, are shown in Table I for each pair (m, λ) appearing in these figures. This table also shows the mean and the variance for each distribution.] For a given delay, the probability density is least for the lowest load. All these densities appear as not-quite-straight lines, slightly convex, in semi-logarithmic plots. The departure from straightness shows the nonexponential character of these functions, which stems from the nonlinearity of the problem. The uppermost curve in Fig. 7, which corresponds to $\Lambda(2)$, appears to have the greatest curvature near $t = 0$, as the phase-plane geometry suggests. The long straight tails in Figs. 7–9 show how good the exponential approximation is for large t.

The delay densities for $\lambda = 0.10102$ and $m = 1.25$, 2, and 3 appear in Fig. 8. For the uppermost curve, $\lambda = \Lambda(3)$, the critical load.

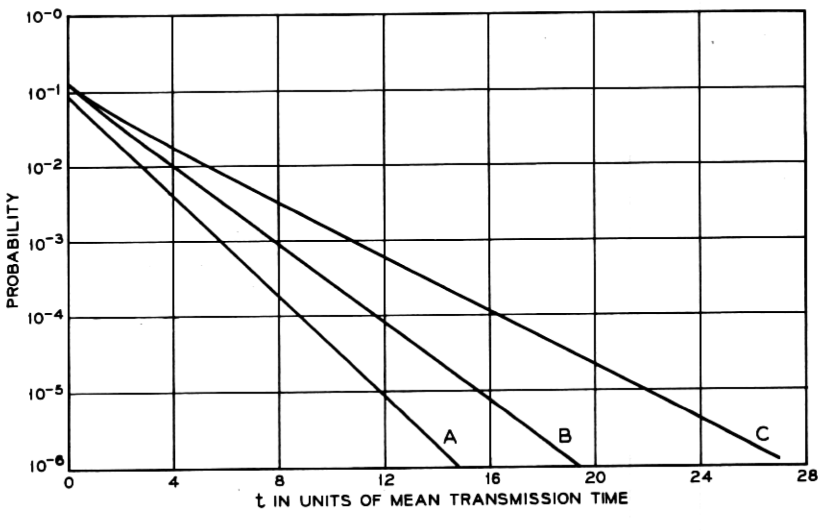

Fig. 7—Delay densities for $m = 2$ and λ equal to (A) 0.10102, (B) 0.15, (C) 0.17157.

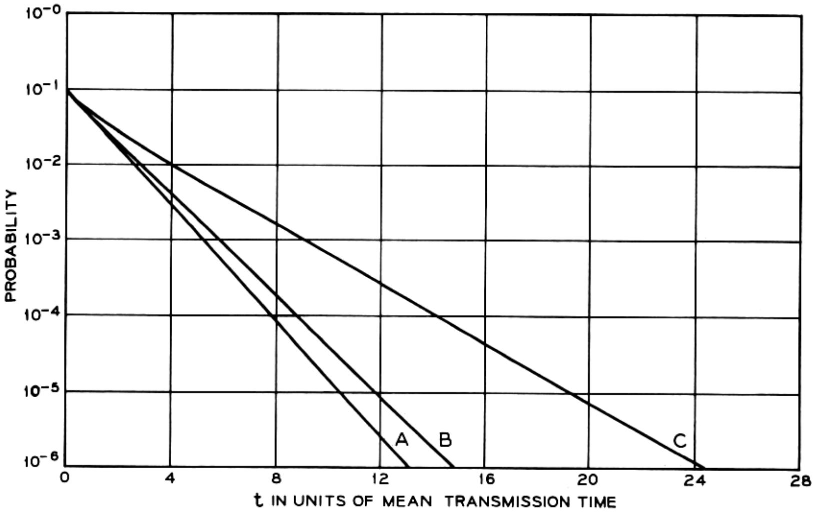

Fig. 8—Delay densities for $\lambda = 0.10102$ and m equal to (A) 1.25, (B) 2, (C) 3.

These functions are shown in Fig. 9 for $\lambda = \Lambda(m)$ and $m = 1.05$, 1.25, 2, 3, and 10. We see that the probability of a given delay *decreases* as m rises; this effect is due of course to the very rapid decrease in critical load with increasing m. In other words, for a large number of addresses per message, equilibrium requires such a small value of offered load that long delays are not likely to occur!

Table I shows that as λ increases for fixed m, or as m increases for fixed λ, F_0 goes down, the mean and variance of the receiver delays increase, and the coefficient of variation goes down. As m increases and λ is kept at its critical value $\Lambda(m)$, all these effects are reversed! However, it is in some ways more instructive to examine the *conditional* coefficient of variation, C_c in Table I: the coefficient of variation of the delay density function, conditional upon the delay being positive. It is C_c that really measures the departure from straightness of the density plots in Figs. 7–9, and we see that these densities become less exponential as λ increases for fixed m, or m increases for fixed λ, or as m *decreases* when $\lambda = \Lambda(m)$.

If the phase-plane trajectory $p(F)$ were straight and the distribution F_λ^T exponential, all the quantities of interest could be calculated exactly without numerical integration. Since C_c is never very far from 1 in Table I, we replace $p(F)$ by a straight line of slope γ_1 and obtain the following approximations for the receiver occupancy and the moments

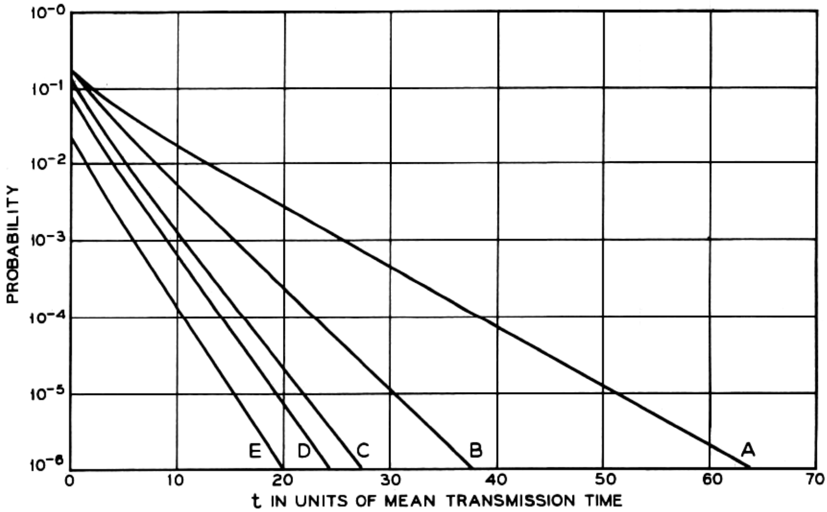

Fig. 9—Delay densities for critical loads and m equal to (A) 1.05, (B) 1.25, (C) 2, (D) 3, (E) 10.

of delay: The line $p = -\gamma_1(1 - F)$ intersects the edge $p = \lambda F$ at the abscissa

$$\tilde{F}_0 = -\gamma_1/(\lambda - \gamma_1). \tag{48a}$$

Likewise we can substitute $-1/\gamma_1$ for $(1 - F)/p(F)$ in (42). When $i = 1$ this yields the estimate $M_1(F) \cong (F_0 - 1)/\gamma_1$, or using (48a),

$$\tilde{E}(\omega) = -\lambda/[\gamma_1(\lambda - \gamma_1)]. \tag{48b}$$

When $i = 2$ it is easier to use (43) directly with $T = 0$, so that $M_2(F) \cong (2/\gamma_1^2)(1 - F_0)$. Substitution of (48a) then gives us

$$\tilde{E}(\omega^2) = 2\lambda/[\gamma_1^2(\lambda - \gamma_1)]. \tag{48c}$$

The estimate (48a) is quite accurate when λ lies well below $\Lambda(m)$. $\tilde{F}_0$ is always too small, but the error does not exceed 1 percent for the three cases in which $\lambda < \Lambda$ in Table I; and for the case $m = 1.25$, $\lambda = 0.10102$ it is only 0.01 percent. For critical load, the error is 19 percent when $m = 1.05$ and $\lambda = 0.64174$, and it decreases to less than 1 percent for $m = 10$.

The error of eq. (48b) is virtually constant when $\lambda = \Lambda(m)$, ranging from 24.5 percent at $m = 1.05$ to 27.3 percent at $m = 10$. $\tilde{E}(\omega)$ is always too large. Like $\tilde{F}_0$, it is a better approximation when $\lambda < \Lambda$, being

TABLE I—PROPERTIES OF RECEIVER DELAY DISTRIBUTIONS AS FUNCTIONS OF LOAD AND MULTIPLICITY

Receiver Load, λ	Mean Number of Addresses per Message, m				
	1.05	1.25	2	3	10
0.64174	$F_0=0.27069$ $C_c=1.075$ $E(\omega)=3.50495$ $C=1.40$ $\mathrm{Var}(\omega)=24.0415$ $\lambda/\Lambda=1$				
0.38197		$F_0=0.50492$ $C_c=1.061$ $E(\omega)=1.41582$ $C=1.82$ $\mathrm{Var}(\omega)=6.6050$ $\lambda/\Lambda=1$			
0.17157			$F_0=0.74481$ $C_c=1.050$ $E(\omega)=0.55612$ $C=2.69$ $\mathrm{Var}(\omega)=2.2383$ $\lambda/\Lambda=1$		
0.15			$F_0=0.80668$ $C_c=1.012$ $E(\omega)=0.31320$ $C=3.08$ $\mathrm{Var}(\omega)=0.9332$ $\lambda/\Lambda=0.8743$		
0.10102		$F_0=0.89606$ $C_c=1.000$ $E(\omega)=0.11941$ $C=4.27$ $\mathrm{Var}(\omega)=0.2602$ $\lambda/\Lambda=0.2645$	$F_0=0.88476$ $C_c=1.003$ $E(\omega)=0.14938$ $C=4.05$ $\mathrm{Var}(\omega)=0.3660$ $\lambda/\Lambda=0.5888$	$F_0=0.84141$ $C_c=1.046$ $E(\omega)=0.32067$ $C=3.49$ $\mathrm{Var}(\omega)=1.2554$ $\lambda/\Lambda=1$	
0.02633					$F_0=0.95599$ $C_c=1.043$ $E(\omega)=0.08281$ $C=6.81$ $\mathrm{Var}(\omega)=0.3184$ $\lambda/\Lambda=1$

off by 6 percent when $m = 2$, $\lambda = 0.15$ and by only 0.15 percent when $m = 1.25$, $\lambda = 0.10102$, and $\lambda/\Lambda = 0.2645$.

The variance of ω, as estimated from (48b) and (48c), is always on the high side and generally less accurate than the approximate mean. The variance is overestimated by 23.5 percent when $m = 1.05$ and by 32.5 percent when $m = 10$, for $\lambda = \Lambda$. The error is 7 percent for $m = 2$, $\lambda = 0.15$, and 0.17 percent for the best case in Table I–that is, when $m = 1.25$ and $\lambda = 0.10102$.

5.4 *Behavior of Systems with Finitely Many Receivers*

In a physical system with fixed finite R, the extent of interdependence among the delays $\omega^{(i)}$ suffered by the copies of one message must increase with λ, as noted on p. 1371. Thus the true distribution of $\omega + \xi$ is less well approximated by F^m as λ grows, and m must be replaced in (26a) by some function of λ which departs increasingly from m. [*Cf.* p. 1404, where $\Delta'(1)$ replaces m.] The critical load for such a system may therefore differ substantially from $\Lambda(m)$.

The predictions of the present theory have been tested against the behavior of physically realizable systems in a modest series of simulation experiments. A few runs were first made to compare the behavior of receiver queues with Poisson arrivals to that found with transmitters in the system. Since there was no perceptible difference with $R = 50$, the remaining experiments simulated only the receiver queues. Arrivals were Poisson; all messages had exactly 2 addresses; and most runs were made with 50 or 200 receivers, although there were several with $R = 100$ and one each with $R = 400$, 500, and 1000.

Representative results are shown in Fig. 10, in which a quantity called "lim F_0" is plotted against load for $m = 2$. The ordinate "lim F_0" represents the asymptotic probability, approached as $t \to \infty$, of finding a receiver idle. In the steady state this quantity is just F_0; and for loads above the critical value, for which the queues grow without bound, it is zero. The theory described in this paper (for infinite R) predicts the lower curve in Fig. 10, which is discontinuous at $\Lambda(2) \cong 0.172$. In a system with exactly m receivers (here $R = m = 2$), all messages go to all m receivers, and the m queues behave as identical copies of a single-server queue. In this case, the critical load is unity; $F_0 = 1 - \lambda$; and "lim F_0" is the linear function, reaching zero at $\lambda = 1$, which is shown as the upper curve in Fig. 10. In this case, the dependence among the receiver queues is complete, and it raises the critical load from $\Lambda(2)$ to 1. The intermediate curves are for $R = 50$ and 200 as shown, and the isolated cross marks an approximate (because of the

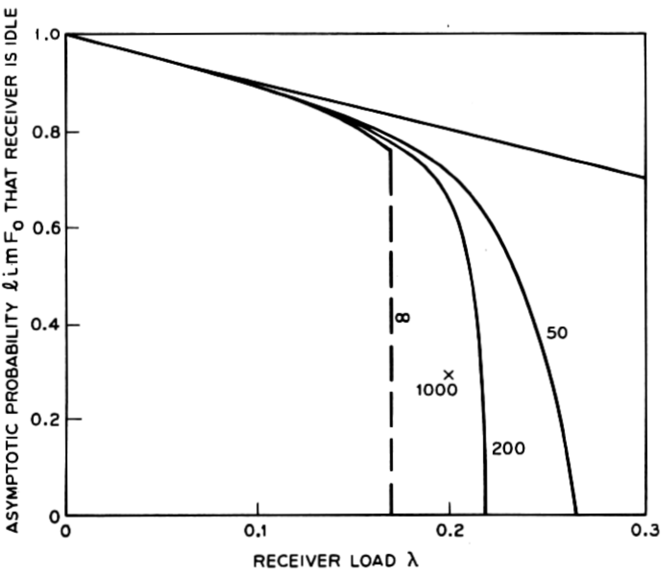

Fig. 10—Idle capacity of receivers as a function of load.

expense of simulation in this case) value for $R = 1000$. We see that the critical load, identified for finite R as the point where "lim F_0" reaches zero, approaches $\Lambda(m)$ from above as R increases. The curves for finite R are shown as slightly above that for infinite R in the region just to the left of $\Lambda(m)$; the difference is small, and the simulated values are not so precise as to guarantee its existence.

Figure 10 indicates that the asymptotic theory is extremely accurate for loads well below Λ and for practical values of R–say from 50 to a few hundred. For loads just below Λ, the behavior of finite systems rapidly approaches that of the infinite model: Even for $R = 50$, the relative error in "lim F_0" at Λ is at most a few percent. The discontinuity in "lim F_0" is not physically realizable: For each finite R, "lim F_0" decreases quickly but smoothly to zero as λ approaches its critical value from below. The family of response curves appears to lie between the straight line for $R = m$ and the discontinuous curve for $R = \infty$ and to approach the latter from the right as $R \to \infty$. This approach is clearly quite slow for $\lambda > \Lambda$, so that the true critical load lies significantly above Λ until R reaches a value at least several thousand. The evidence is similar for the other parameters measured in the simulations: $E(\omega)$, $E(\omega + \xi)$, mean queue-length, and occupancy.

More than this we cannot learn by experiment, at least at acceptable cost.

In summary, the evidence from simulation indicates that our model represents reality well in the range $\lambda \leqq \Lambda$, and that for $\lambda > \Lambda$ it is valid asymptotically but with very slow convergence in R. A precise analysis of the effects of dependence for finite R remains an interesting and difficult question for further research.

VI. TRANSMITTER DELAYS

We now return to the question of transmitter delays which was left behind on p. 1373. The distribution of delay δ could be found numerically from the results of the previous section by evaluating the equilibrium form of (3),

$$G(t) = e^{-t} \int_0^t e^u F^m(u)\, du,$$

to find the distribution G of transmitter service-times ρ; calculating numerically the Laplace transform of G; and numerically inverting the transform, which is obtained from Pollaczek's formula, of the δ-distribution. Such extensive computations do not seem justified in a study of the present kind, and we consider here only the mean delay.

We recall from Fig. 2 and p. 1369 that the receiver sojourn-time is also $\rho = \omega + \xi + x$. The receiver queue can be viewed as a single-server system with delay ω and service time $\xi + x$. Its occupancy is therefore $\lambda E(\xi + x)$ and can also be written $1 - F_0$. Since $\xi + x = \rho - \omega$, we have $\lambda E(\rho - \omega) = 1 - F_0$, so that

$$E(\rho) = E(\omega) + \frac{1 - F_0}{\lambda}. \tag{49}$$

The total receiver-delay suffered by the average *message*, through having to wait in m queues before transmission, is $E(\omega + \xi) = E(\rho - x) = E(\rho) - 1$, or

$$E(\omega + \xi) = E(\omega) - 1 + \lambda^{-1}(1 - F_0). \tag{50}$$

We call this the *receiver queuing-time.*

In order to evaluate $E(\delta)$ by means of the Pollaczek-Khinchin formula, we need two more quantities, of which the first is the transmitter occupancy. Recalling from (1) that the transmitter arrival-rate is $\alpha = k\lambda/m$, where k is the ratio R/X of the numbers of receivers and transmitters, we can write the transmitter occupancy as

$$\alpha E(\rho) = (k/m)[1 - F_0 + \lambda E(\omega)] \tag{51}$$

from (49). The second quantity required is $E(\rho^2)$, the second moment of transmitter service-times. Differentiation of the equilibrium form of (10) yields $\lambda \dot{G} = \lambda \dot{F} - \ddot{F}$, which, after multiplication by t^2 and integration from zero to infinity, becomes

$$\lambda E(\rho^2) = \lambda E(\omega^2) - \int_0^\infty t^2 \ddot{F}(t)\, dt.$$

We integrate by parts and divide through by λ to obtain

$$E(\rho^2) = E(\omega^2) + (2/\lambda)E(\omega). \tag{52}$$

The Pollaczek-Khinchin formula (Ref. 2, p. 117) for the mean transmitter-delay involves the transmitter occupancy and the first and second moments of its service time ρ:

$$E(\delta) = \frac{\alpha E(\rho)}{2[1 - \alpha E(\rho)]} \cdot \frac{E(\rho^2)}{E(\rho)}.$$

If we cancel the factors $E(\rho)$ and substitute into the resulting equation the relations (1), (51), and (52), and simplify, we find that

$$E(\delta) = \frac{E(\omega) + (\lambda/2)E(\omega^2)}{(m/k) - [1 - F_0 + \lambda E(\omega)]}. \tag{53}$$

Adding (50) and (53), we obtain the total delay suffered by the average message, prior to transmission, in both transmitter and receiver queues.

We are now in a position to consider the choice of k. The simplest approach is to let $k = m$, which results in equal *utilization* of receivers and transmitters; that is, all terminals spend the same fraction of time with transmission of messages actually taking place. In this case, (53) reduces to

$$E_u(\delta) = \frac{E(\omega) + (\lambda/2)E(\omega^2)}{F_0 - \lambda E(\omega)}. \tag{53a}$$

The denominator of this expression need not be positive. From (49), $F_0 - \lambda E(\omega) = 1 - \lambda E(\rho)$, and the last term is the average number of messages present (waiting and being served) at a single receiver, which need not be less than 1. For fixed m and k, the transmitter delay becomes infinite as the load (α and λ) increases, and this may occur at a value $\lambda < \Lambda(m)$. In other words, for fixed m and $\lambda < \Lambda(m)$, it may be impossible to utilize transmitters as efficiently as receivers: The ratio k may have to be smaller than m. As shown in Table II below, this phenomenon occurs only for m quite close to 1, so that $\Lambda(m)$ is relatively large.

We could also choose the ratio of numbers of receivers to transmitters by making their *occupancies* equal; that is, all terminals spend the same fraction of time with at least one message present (waiting, camping on, or being transmitted). The condition for this is that $1 - F_0 = \alpha E(\rho)$, which from (51) becomes

$$k = m(1 - F_0)/[1 - F_0 + \lambda E(\omega)]. \tag{54}$$

The corresponding expected transmitter-delay is

$$E_o(\delta) = \frac{(1 - F_0)[E(\omega) + (\lambda/2)E(\omega^2)]}{F_0[1 - F_0 + \lambda E(\omega)]}. \tag{53b}$$

As a third alternative, we could require the mean delays suffered by a message at the transmitter and receiver stages to be equal. The condition for this is that $E(\omega + \xi) = E(\delta)$, which from (50) and (53) yields

$$\frac{m}{k} = 1 - F_0 + \lambda E(\omega) + \frac{\lambda E(\omega) + (\lambda^2/2)E(\omega^2)}{1 - F_0 - \lambda + \lambda E(\omega)}. \tag{55}$$

In this case $E(\delta)$ is given simply by (50).

In a practical context, m and α would be given, and k (and hence λ) would be determined by an appropriate balance of hardware and delay costs. An optimum design can be found only in terms of receiver costs, transmitter costs, and a cost per unit of delay time. We do not pursue this (still oversimplified) system-design problem, but merely illustrate relative performance of transmitters and receivers in terms of the criteria mentioned above. This is done in Table II, which was constructed by calculating the receiver-transmitter ratio k, the transmitter arrival-rate α, the mean transmitter-delay $E(\delta)$, and the mean total delay for a message "Del" = $E(\delta + \omega + \xi)$. These quantities are shown for each design-criterion mentioned above and for each combination of mean number of addresses m and receiver arrival-rate λ listed in Table I. For each choice of m and λ, the mean receiver queuing-time per message $E(\omega + \xi)$, which is not affected by the choice of k, is also shown. This quantity comes from (50); the others come from (54), (55), (1), (53a), and (53b). "Del" is the sum of $E(\delta)$ and $E(\omega + \xi)$.

In the two configurations for which $F_0 - \lambda E(\omega) < 0$, $E(\delta)$ is not defined; it becomes infinite at some smaller value of α. We indicate in Table II how far beyond this singularity α lies by giving the ratio $F_0/(\lambda E(\omega))$. The locus of this singularity could be determined, but is

TABLE II—EXPECTED TRANSMITTER DELAY AND TOTAL DELAY FOR VARIOUS SYSTEM CONFIGURATIONS

	Criterion for Choice of $k = R/X$		
Case	Equal Utilization	Equal Occupancy	Equal Delay
$m = 1.05, \lambda = 0.64174$ $E(\omega + \xi) = 3.641$	$k = 1.05$ $E(\delta) = \infty$ $\alpha = 0.642$ $[F_0/\lambda E(\omega) = 0.120]$	$k = 0.257$ $E(\delta) = 13.71$ $\alpha = 0.157$ Del $= 17.35$	$k = 0.147$ $E(\delta) = 3.641$ $\alpha = 0.090$ Del $= 7.282$
$m = 1.25, \lambda = 0.38197$ $E(\omega + \xi) = 1.712$	$k = 1.25$ $E(\delta) = \infty$ $\alpha = 0.382$ $[F_0/\lambda E(\omega) = 0.934]$	$k = 0.597$ $E(\delta) = 2.897$ $\alpha = 0.183$ Del $= 4.609$	$k = 0.443$ $E(\delta) = 1.712$ $\alpha = 0.135$ Del $= 3.424$
$m = 2, \lambda = 0.17157$ $E(\omega + \xi) = 1.044$	$k = 2$ $E(\delta) = 1.193$ $\alpha = 0.172$ Del $= 2.237$	$k = 1.46$ $E(\delta) = 0.757$ $\alpha = 0.125$ Del $= 1.801$	$k = 1.83$ $E(\delta) = 1.044$ $\alpha = 0.157$ Del $= 2.088$
$m = 2, \lambda = 0.15$ $E(\omega + \xi) = 0.596$	$k = 2$ $E(\delta) = 0.514$ $\alpha = 0.15$ Del $= 1.110$	$k = 1.61$ $E(\delta) = 0.389$ $\alpha = 0.121$ Del $= 0.985$	$k = 2.24$ $E(\delta) = 0.597$ $\alpha = 0.168$ Del $= 1.193$
$m = 1.25, \lambda = 0.10102$ $E(\omega + \xi) = 0.148$	$k = 1.25$ $E(\delta) = 0.151$ $\alpha = 0.101$ Del $= 0.299$	$k = 1.12$ $E(\delta) = 0.133$ $\alpha = 0.091$ Del $= 0.281$	$k = 1.23$ $E(\delta) = 0.148$ $\alpha = 0.100$ Del $= 0.296$
$m = 2, \lambda = 0.10102$ $E(\omega + \xi) = 0.290$	$k = 2$ $E(\delta) = 0.194$ $\alpha = 0.101$ Del $= 0.484$	$k = 1.77$ $E(\delta) = 0.169$ $\alpha = 0.089$ Del $= 0.459$	$k = 2.81$ $E(\delta) = 0.290$ $\alpha = 0.142$ Del $= 0.580$
$m = 3, \lambda = 0.10102$ $E(\omega + \xi) = 0.891$	$k = 3$ $E(\delta) = 0.481$ $\alpha = 0.101$ Del $= 1.372$	$k = 2.49$ $E(\delta) = 0.384$ $\alpha = 0.084$ Del $= 1.275$	$k = 4.78$ $E(\delta) = 0.891$ $\alpha = 0.161$ Del $= 1.782$
$m = 10, \lambda = 0.02633$ $E(\omega + \xi) = 0.754$	$k = 10$ $E(\delta) = 0.091$ $\alpha = 0.0263$ Del $= 0.845$	$k = 9.53$ $E(\delta) = 0.087$ $\alpha = 0.025$ Del $= 0.841$	$k = 61.8$ $E(\delta) = 0.754$ $\alpha = 0.163$ Del $= 1.508$

not of sufficient theoretical interest to calculate here. Suffice it to observe that the delays are in some sense dominated by the transmitter queues for small m and by the receiver queues for large m: For $m = 1.05$ and $\lambda = \Lambda$, transmitter arrival-rates must be much smaller than λ in order for the transmitter queues to be in equilibrium at all; for m greater than some value not much above 1.25, k can equal m for any permissible λ; and for $m = 10$, a very small number of transmitters can handle, without excessive delay, all the traffic that can pass through the receiver queues. This phenomenon is another consequence of the extent to which the camp-on discipline limits receiver utilization at high address-multiplicities. The latter effect is partly illustrated by comparing the mean time spent camping on, $E(\xi)$, with the ordinary receiver-delay $E(\omega)$: The ratio $E(\omega + \xi)/E(\omega)$ generally increases with m, ranging from 1.04 for the first case listed in Table II to 9.1 for the last, and being always a little smaller than m.

The approximation based on assuming F_λ^T to be exponential, with decay-constant $-\gamma_1$, can be used to estimate the quantities of interest in this section as it was in the last. Direct substitution of eqs. (48) yields these "linear" estimates in terms of λ and γ_1 alone: First, from (50),

$$\tilde{E}(\omega + \xi) = -(1 + \gamma_1)/\gamma_1 . \tag{56a}$$

This estimate is always on the high side. Its error is largest in the eighth case in Table II: about 40 percent for $m = 10$, $\lambda = \Lambda$. Next, for the criterion of equal utilization, (53a) leads to

$$\tilde{E}_u(\delta) = \lambda/[\gamma_1(\lambda + \gamma_1)]. \tag{56b}$$

This approximation too is always high, and is worst, with an error of 43 percent, in the third case of Table II, with $m = 2$ and load Λ. [Of course it must get still worse near the singularity of $E_u(\delta)$, which (56b) estimates wrongly as occurring where $\lambda = -\gamma_1$.] For the criterion of equal occupancy we get

$$\tilde{k}_o = m\tilde{F}_0 \tag{56c}$$

from (54) and

$$\tilde{E}_o(\delta) = \lambda/\gamma_1^2 \tag{56d}$$

from (53b). The value $\tilde{k}_o$ is too low in the eight examples considered here, the largest error being about 11 percent at $m = 1.05$; and $\tilde{E}_o(\delta)$ is too large in these examples, with a maximum error of 46 percent at $m = 1.05$. Last, (55) becomes

$$\tilde{k}_d = -m\gamma_1(1 + \gamma_1)/[\lambda(2 + \gamma_1)], \tag{56e}$$

which corresponds to the criterion of equal delay. The error of this estimate takes both signs in these examples, and is worst, at about -10 percent, when $m = 1.05$.

No rigorous analysis of the errors of these approximations has been performed. However, all the estimates in eqs. (48) and (56) are better for $\lambda < \Lambda(m)$ than for critical load; and the cases reported in Tables I and II include a representative sample of values of m from 1.05 to 10, for $\lambda = \Lambda$. Furthermore, all the errors behave monotonically in m for $\lambda = \Lambda(m)$. Since the aforementioned range of m covers all the values that seem likely to arise in communications engineering, it is reasonable to conclude that the estimates (48) and (56), with γ_1 known exactly from (25), can be used in place of exact results of the asymptotic theory whenever a maximum relative error of 50 percent–larger than any of those encountered in these examples–is tolerable in the tabulated quantities.

The present theory predicts no startling qualitative behavior of the transmitter queues, as it does for the receivers. As noted in Section 5.4, the receiver queues behaved no differently in the presence of transmitters than they did with Poisson input, in the few cases simulated (with both sub- and supercritical loads). Thus no detailed records were kept of the transmitter-queue parameters in these simulation runs; and the remarks of Section 5.4 may be considered to apply to the whole system as well as to the receivers alone.

VII. DISCUSSION

This section includes a brief summary of the argument and results of this paper, a discussion of its relation to other literature, and a statement of problems that remain open.

7.1 *Summary*

After an Introduction relating the camp-on problem to the question of engineering for multiple-address traffic in data communication systems, a specific model of a camp-on system is described in Section II. This model is reduced by an informal argument to an idealized mathematical model of receiver delays, of interest in its own right and characterized by Poisson arrivals, exponential transmission-times, and the fact that each message can be transmitted only after the longest of m independent delays in receiver queues has ended. [Key symbols appear in eq. (1) and Fig. 2.] The mathematical model yields the integral recurrence (7) for the distributions of successive delays at a given receiver. This recurrence leads to an integral equation [eq. (12)] in statistical equilibrium and also to an equivalent differential system [eqs. (18), given in Section III]. The method used to establish the fundamental recurrence, though based on Lindley's equation, is new in its treatment of the delays suffered by a message in parallel queues, and very simple; it should prove useful in other traffic problems involving messages which must wait in several queues at once. Another important element of our approach is a method for separation of the transmitter and receiver delay-analyses, which should be useful in other two-stage queuing problems in which servers in both stages are released (or seized) simultaneously.

In Section III the equation for the equilibrium distribution of receiver delays is reduced to the first-order differential system (20). Analysis of the corresponding vector-field and its topology in the phase plane shows that, for receiver arrival-rates λ not exceeding a critical load Λ, a one-parameter family of distributions exists, each of which satisfies all requirements for a solution to the problem of delays encountered in equilibrium. The uppermost member of this family is qualitatively distinct from all the others. The critical load Λ is found explicitly [eq. (26a)] as a function of the number m of addresses per message; above this load, statistical equilibrium cannot exist.

In Section IV we return to the recurrence (7) to show that, although the integral equation (12) has infinitely many solutions, any reasonable assumption about the previous history of the system leads to a unique limiting-distribution when the load λ is held constant indefinitely at a value smaller than Λ; and the distribution in question is the uppermost of the equilibrium solutions. A slightly weaker result holds when $\lambda = \Lambda$: The existence of the limit can be guaranteed only if the system has not previously been subjected to too great an overload. Even if it has, delays do not increase indefinitely when the load is held at its critical value; instead, the distributions of delay encountered by all subsequent messages are bounded below by some member of the family of stationary distributions corresponding to $\lambda = \Lambda$. (See Theorems 4 and 5 and Corollaries 3 and 4.) So far as I am aware, no other example has been reported of a queuing system which can operate in equilibrium–and with delays having finite moments of all orders–*at* (not just *below*) its critical load when the basic service-process (here exponential) admits of arbitrarily long holding-times. This startling result applies, of course, only to the asymptotically large system which is not physically realizable. The structure of the proofs in Section IV also shows that analytical techniques are necessary, and that numerical study of the sequences generated by (7) could not solve the uniqueness problem.

A procedure for calculating the limiting distributions of receiver delay is described in Section V. The computation of the phase-plane trajectories near a singular point required considerable effort, and when λ was near Λ rested on the use of a generally neglected series expansion developed by Briot and Bouquet in 1856 and by Picard in 1908.[7] Graphs of eight typical distributions are shown (Figs. 7–9) along with their means, variances, and probabilities of no delay (Table I). The receivers spend so much time being camped on, and this factitious loading so limits their useful capacity, that mean delays and probabilities of delay generally *decrease* as the number m of addresses per message rises and λ stays at the same fraction of critical load. This remarkable behavior shows how inefficient is the camp-on discipline in its pure form. In Section V, it is also shown that the previous analysis can be validly interpreted with minor quantitative changes when m is not an integer. A consequence of this model is the explicit representation of $\Lambda(m)$, which is shown (*cf.* Fig. 6) to decrease so rapidly above $m = 1$ as to account for the curious reduction of delays mentioned above.

The numerical predictions of the asymptotic theory are compared with simulation results in Section 5.4. For a physical system with a finite number R of receivers, the true critical load, along with other indicators of performance such as the probability F_0 of finding a receiver idle, depends on R. Convergence to the predicted behavior as R increases is rapid for $\lambda \leqq \Lambda(m)$, so that in this range the idealized model is very accurate for such values of R as are likely to arise in engineering. When $\lambda > \Lambda(m)$, the effects of interdependence among the receiver queues dissipate very slowly as R approaches infinity. For fixed finite R, the critical load exceeds $\Lambda(m)$ and the idle capacity of the receivers tends to zero as the load increases toward its critical value. The critical load falls with increasing R, approaching $\Lambda(m)$ from above, and the changes in performance parameters that occur as λ approaches its critical value from below become more abrupt. The discontinuity in system behavior at $\Lambda(m)$, that is characteristic of the idealized system, cannot be realized and is approached only asymptotically by response curves for increasing R. Fig. 10 illustrates these effects.

In Section VI the earlier results on receiver performance in the asymptotic model are related to the behavior of the transmitter queues. The choice of the design parameter k is considered (k being the ratio of numbers of receivers to transmitters), and the mean transmitter-delay is calculated for some representative configurations (Table II).

Each transmitter queue behaves as an ordinary M/G/1 system whose service time is the receiver sojourn-time. Critical receiver-loads are so strictly limited for large m that queuing at transmitters is not a severe problem; but when m is near 1, receiver delays contribute so greatly to transmitter occupancies that transmitter arrival rates must be held substantially below those for receivers.

7.2 *Related Literature*

Many of the most important unsolved problems of congestion theory relate to queues in parallel or in series which interact in complex ways. The work reported here is significant partly because it constitutes a partially successful attack on a problem of this type and may, as discussed above, lead to the solution of others. An early example of studies of the same class is the elegant paper by Kingman[9] on two queues in parallel, where each new arrival joins the shorter queue. Studies of other important mechanisms of interaction are gradually becoming more common in the literature. I cite as a recent example involving communications traffic the paper by Cooper and Murray.[10] Hunter[11] surveys the literature on exactly two queues in parallel; and a heavy-traffic approximation for many such queues, with customers randomly choosing which queue to join, is given by Whitt.[12]

It was mentioned in Section I that the camp-on problem itself is treated in a paper by Haenschke.[1] He does not consider the delays encountered at transmitters, or the relation between the transmitting and receiving stages of a camp-on system, but analyzes the receiver queues–the essential component of the problem–by means of a clever approximation. He assumes, as we do, that arrivals at each receiver are Poisson, that transmission times are exponential, and that delays in different receiver queues are independent. He also assumes that all receiving locations have the same number of lines; in our model this number is always 1, but in Haenschke's paper it can be any positive integer. His technique is based on the assumption that the receiver service-time $\xi + x$ is exponentially distributed, so that by Erlang's delay theory (see Ref. 2 or Ref. 4, for example) the receiver delay distribution F is also exponential. The resulting model is in essence a linear approximation to the nonlinear one analyzed here. Haenschke's results (which are not directly comparable with those presented in Section V above) are adequately convincing with regard to the drawbacks of the camp-on discipline in practice. They do not, on the other hand, yield any inkling of the qualitative implications of assuming the receiver queues to behave independently.

This system is also discussed by Weber[13] in an unpublished report which includes extensive simulation results. At loads below our values of $\Lambda(m)$, where the magnitude of R is not very important, his figures agree closely with those presented here. For example, when $m = 2$ and $\lambda = 0.102$, Weber's simulation gives for the receiver occupancy and the mean queuing-time per message [$1 - F_0$ and $E(\omega + \xi)$ in our notation] the values 0.116 and 0.287. Our calculation for $\lambda = 0.101$ (from Tables I and II) yields 0.115 and 0.290 respectively (and see Fig. 10). There being no extant record of the values chosen for R in Weber's runs, his results cannot be compared in detail with ours when $\lambda > \Lambda(m)$. However, examination of his printout shows no inconsistency with the discussion in Section 5.4 above.

It is amusing to observe in closing that our basic differential equation (18a) agrees, except for the sign of the coefficient of F, with the homogeneous (zero driving-term) equation for the anharmonic oscillator. This equation has recently been studied (see Bloembergen,[14] for example) in connection with nonlinear optics. However, the necessary methods do not overlap: We are interested in decaying solutions, while in optics the oscillatory solutions (*cf.* the change of sign just mentioned) are relevant and are obtained by perturbation techniques good only for very small values of the parameter we call λ.

7.3 *Open Questions*

A number of issues raised in this paper are clearly in need of further investigation. Most important, of course, would be an exact analysis of the receiver queues for finite R, quantifying the effects of dependence among them and the rate of convergence to the asymptotic model as $R \to \infty$. It would be particularly useful and interesting to have an analytical expression for the critical load as a function of R as well as m.

Our results should be extended to cover other arrival and transmission-time distributions, especially the case of constant message-length. It would be important to solve the present problem (and many other queuing problems!) without the assumption of complete symmetry–that is, of equal loads on all receiving stations. And the present work should of course be extended to the case of more than one receiving line per location, which was treated by both Weber[13] and Haenschke.[1]

This paper reports on a technique which is new in detail, though not in principle, and describes a curious qualitative result on the behavior of a queuing system (albeit not a physically realizable one) at critical load. It will be interesting to test the technique on other problems involving complex interactions between queues, and to find the domain of validity of the qualitative result.

VIII. ACKNOWLEDGMENTS

I am grateful to a number of colleagues for their help: to R. W. Hamming, J. A. Morrison, and especially David R. Cox for stimulating discussions and useful advice; to A. Kuczura for acquainting me with Picard's method, which made the computations possible in the case of critical load; to Barbara R. LaCava for obtaining the simulation results for finite R and for her diligent pursuit of the delay distributions; to P. J. Burke, A. Descloux, E. O. Elliott, A. J. Fabens, M. Segal, and the referee for useful comments on various drafts of this paper; and most of all to Henry J. Landau for much help, for the idea which led to the results of Section IV, for penetrating comments on an early draft of this paper, and for constant encouragement, without which this work would not have been completed.

REFERENCES

1. Haenschke, D. G., "Analysis of Delay in Mathematical Switching Models for Data Systems," Bell Sys. Tech. J., *42* (1963), pp. 709–736.
2. Khintchine, A. Y., *Mathematical Methods in the Theory of Queueing*, 2nd ed., London: Griffin, and New York: Hafner, 1969. (First published in Russian, 1955.)
3. Cox, D. R., and Smith, Walter L., "On the Superposition of Renewal Processes," Biometrika, *41* (1954), pp. 91–99.
4. Cox, D. R., and Smith, Walter L., *Queues*, London: Methuen, and New York: John Wiley & Sons, 1961.
5. Lindley, D. V., "The Theory of Queues with a Single Server," Proc. Camb. Phil. Soc., *48* (1952), pp. 277–289.
6. Coddington, E. A., and Levinson, N., *Theory of Ordinary Differential Equations*, New York: McGraw-Hill, 1955.
7. Picard, Émile, *Traité d'Analyse*, 3rd ed., Paris: Gauthier-Villars, 1928, Vol. 3, Chap. II, Sec. II, pp. 30–34.
8. Kaplan, W., *Advanced Calculus*, Reading: Addison-Wesley, 1952.
9. Kingman, J. F. C., "Two Similar Queues in Parallel," Ann. Math. Statist., *32* (1961), pp. 1314–1323.
10. Cooper, R. B., and Murray, G., "Queues Served in Cyclic Order," Bell Sys. Tech. J., *48* (1969), pp. 675–689.
11. Hunter, Jeffrey J., "Two Queues in Parallel," J. Roy. Statist. Soc. (**B**), *31* (1969), pp. 432–445.
12. Whitt, Ward, "Multiple Channel Queues in Heavy Traffic. III: Random Server Selection," Adv. Appl. Prob., *2* (1970), pp. 370–375.
13. Weber, J. H., "Delays in Line Switching Systems with Multiple Address Traffic," unpublished work (1959).
14. Bloembergen, N., *Nonlinear Optics*, New York: Benjamin, 1965.