CONNECT TIMES OF SWITCHED DATA NETWORKS WITH DELAY

S. Bing Han

A DISSERTATION

in

The Faculty

of

Engineering

Presented in Partial Fulfilment of the Requirements for

the Degree of Master of Engineering at

Sir George Williams University

Montreal, Canada

April, 1973

ACKNOWLEDGMENTS

I would like to thank Bell-Northern Research for their permission to use this report as a thesis to be submitted to Sir George Williams University in partial fulfilment of the requirements for the degree of Master of Engineering.

Thanks are also due to Dr J.C. Giguere, my thesis supervisor, and Dr F.A. Gerard for their encouragement.

SUMMARY

In switched data communications there are basically two methods of transferring data from calling subscriber to called subscriber, namely by circuit-switching technique or by store-and-forward technique (either as message or packet switching). Advantages and disadvantages are associated with both of these switching modes depending on switching times of the data switching machines, the message length, propogation delay as with terrestial vs satellite transmission, security, etc.

Port and Closs[1] compared a circuit-switching and a message-switching system on the basis of waiting times for a one-link and a three-link connection. The delay due to processing time in the switching machines and propagation were neglected.

This report investigates the influence of the delay due to the finite processing time of the switching machine. The effect of propagation delay on the connect time in a circuit and message switching system which is important on communication links using satellite, is also discussed.

Curves plotted with the initial offered traffic as variable shows the maximum loading of the trunk lines and hence give an indication of the "utilization" of the transmission channels.

Packet switching is then introduced, and the resulting waiting times discussed.

Some of the important conclusions of this study are:

a) The connect time for circuit switching is in general smaller than for the message-switching technique for low traffic.

b) For a multiple link connection, the value of the traffic intensity ρ whereby the connect time for circuit switching becomes larger than for message switching, can be increased by increasing the number of channels.

c) In the store-and-forward technique, the connect time can be decreased by introducing packet switching.

c) In store-and-forward technique it is desirable to keep the delays due to the switching machine very small relative to the transmission time of the average message.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

In switched data communications, where a data terminal can, on demand, establish communication with another over a network, two basic techniques are currently considered: circuit switching and store-and-forward switching.

In *circuit switching*, a transmission channel is first set up between the originating and the terminating data terminal. The information is then transmitted virtually instantaneously except for a very small constant delay. The channel between data terminals is maintained during the duration of the call. It may be an identifiable physical path established, for example by means of mechanical contacts in the switching network as in the step-by-step or cross-bar switching (space division switching) or it may be specific time-slots in a system with time division switching. In general, systems are hybrid with some of the shared facility being space, a time slot in the time domain or a frequency band in the frequency domain. Whatever the means that is used to establish the channel, it remains associated with that particular connection during the entire duration of the information exchange.

With *store-and-forward switching* there is no direct transmission path between the two terminals. Two versions of this are message switching and packet switching.

In *message switching*, the originating terminal delivers the complete message and the address of the terminating terminal to the switching machine at the first office, and delegates the further transmission of the message to the switching machine. The latter receives the message, stores it and, from the addressing information, selects the correct routing to the next switching machine or terminating data terminal. When a transmission channel is available, the message is then relayed to the intermediate switching machine (or terminating terminal). At the intermediate switching machine the same tasks take place: reception, storage, and re-transmission at a later time. The message may go through one or more switching machines until it reaches its destination. At each switching machine the time between reception and retransmission is a variable, and depends on the traffic. In addition, retransmission can take place at a different speed than the incoming speed. At each instant in time there is no particular path between the originating and terminating data terminals.

In *packet switching*, a long message cannot be transmitted intact. It is subdivided into short sub-messages (packets), and each packet is provided with a header (i.e., address information). These packets are sent through the network and combined at the receiving node to form the original message. The transmission of each packet proceeds similar to message switching.

A classical example of a message-switching system is the torn-paper tape telegraph, where the punched papertape is torn from the receiving equipment and routed to a transmitter for later transmission to its destination. An electronic equivalent is, for example, the Message Switching Data System of Bell Canada. Packet switching is employed in the ARPA (Advanced Research Project Agency) network in the U.S.A.[10] and in the experimental network of the National Physical Laboratories in the U.K.[11].

REMARKS: If we compare the two methods of message transmission with time division multiplexing, then circuit switching is analogous to synchrounous time-division multiplexing, while store-and-forward switching can be likened to asynchronous time-division multiplexing. Chu[13,14] studied this problem in relation to time-sharing computers, and gave, in his papers, design data on buffer size, probability of overflow, etc.

Some of the characteristics of circuit switching and store-and-forward techniques are[2,12]:

CIRCUIT SWITCHING

a) Transmission rate between users must be the same, otherwise storage is needed at the terminal to buffer the speed difference. This does not imply that transmission rates in both directions have to be equal.

b) This is the most efficient method for transmitting large amounts of data (bulk transmission).

c) No formatting of data is required by the network.

d) Once transmission path is established, long-duration messages are more secure against loss of data.

e) Transmission delay for an established call is fixed.

STORE-AND-FORWARD

a) Bit rate between users can be different; storage is provided by the network and thus differences in speed between originating and terminating can be accepted.

b) This method makes efficient use of transmission facilities, because at each switching center the messages/packets are independently "queued," and "pauses" in data stream are used (e.g., "think" times in conversational computer systems).

c) Length of message is restriced in message switching. Long duration messages have to be partitioned to fit the format required by the network (as in packet-switching).

d) Transmission delay is variable.

In this report we will only compare the two systems on the basis of connect time and utilization of facilities, and will not elaborate on the other advantages and/or disadvantages.

2. DEFINITIONS AND NOTATIONS

2.1 DEFINITIONS OF TRANSFER AND CONNECT TIMES

Consider a *circuit-switching system* with two switching centers S_1 and S_0 (Figure 1). Assume the system is a waiting system (i.e., calls arriving at the switching center S_1 when no circuit is available will be put in a queue until a line becomes free). Assume also that it uses the channel for data as well as for signaling purposes, and hence it does not use centralized control signaling. A call from Data Terminal Equipment (DTE) A for DTE B arrives at switching center S_1 and contends for a free channel to S_0. If DTE B is free, the call will be connected to B and the transmission of the information can begin. Assume that the local data rate from A to S_1 is c_0 b/s (bits per second) and there are N trunks of capacity c_0 b/s each connecting S_1 to S_0 and that the average message length is L bits.

The total transfer time, for which the channel between A and B is fully or partly occupied, consists of:

T_1 = queueing time at S_1. (We adopt the definition: waiting time = total time spent in system = queueing time + service time). T_1 is a function of the number of message arrivals per unit time and the service time at S_1. This latter is the time the channel between S_1 and B is occupied and is equal to $(T_2 + T_3 + T_4 + T_5)$.

T_2 = switching time in S_1 and S_0 plus signaling time between S_1 and S_0, plus signaling time between S_0 and B during the establishment of a connection. Signaling time between S_1 and S_0 will depend on propagation delay and this propagation delay will have to be accounted twice, if confirmation by B is required (assuming full availability in S_0, no queueing at switch S_0 and B not busy).

T_3 = the transmission time of the message itself.

T_4 = propagation delay for the message A to B.

T_5 = disconnect time; if the trunk is used for one-way transmission only, this part can be neglected in calculating the service time of S_1.

Neglecting T_5, the transfer time for the whole message in the circuit switching system is then $(T_1 + T_2 + T_3 + T_4)$.

If the connect time is defined as the time elapsed between call request and first message bit arriving at B, then this is equal to $(T_1 + T_2 + T_4)$. As a first approximation, $T_4 + T_2$ can be lumped

together to form a constant delay T_2 in the system. The connect time will be denoted by τ_c for the circuit-switching system. Thus

$$\tau_c = T_1 + T_2 \qquad \text{(as a first approximation)}.$$

In the *message-switching system* of Figure 2, subscriber A sends its message to S_1 at the local subscriber rate. At S_1 the message is stored, and queued for a free time slot in the high-speed link between S_1 and S_0, then it is transmitted at the channel rate of c to S_0, where the message is again stored. At S_0 the message is retransmitted to its destination B at a speed c_0', not necessarily equal to c_0. The total transfer time then consists of:

T_a = transmission time of message from A to S_1 at transmission rate of c_0 b/s.

T_b = waiting time in the switch S_1, which consists of switching time plus queueing time plus transmission time over trunk at a transmission rate of c b/s.

T_c = assuming again a full availability group, and terminal B to be free, this will be equal to switching time plus transmission time over local line (loop) at a rate c_0'.

T_d = propagation delay from S_1 to S_0.

Let us consider T_c. If the speed c_0' over the local S_0 to B is equal to c_0, and retransmission takes place only after the complete message is received, this term will be equal to T_a plus switching time at S_0. However, the switch S_0 can be designed such that it retransmits the message bit by bit immediately after receipt of B's address, in which case T_c will be small compared to T_a. Alternatively the link between S_0 and B could be a high-speed link, if the DTE B is a computer port; again T_c will be much smaller than T_a. In our further analysis we will therefore neglect T_c. It can always be added later on.

In order to compare the "connect time" of this message-switching system with the circuit-switching system, we define the connect time for the message switching system as

$$\begin{aligned} \tau_m &= \text{transfer time} - T_a \\ &= T_b + T_c + T_d \\ &\approx T_b + T_d \quad \text{(if } T_c \text{ is negligible)}. \end{aligned}$$

Although the propagation delay T_d is included in the connect-time definition, this delay has to be excluded in the calculation of service time of S_1, since messages from S_1 to S_0 can be directly strung together. The switching machine does not have to wait for confirmation from B. Each message is a complete entity with full information as to its destination. This is unlike the circuit-switching system, where the propagation delay enters the equation for the service time of S_1.

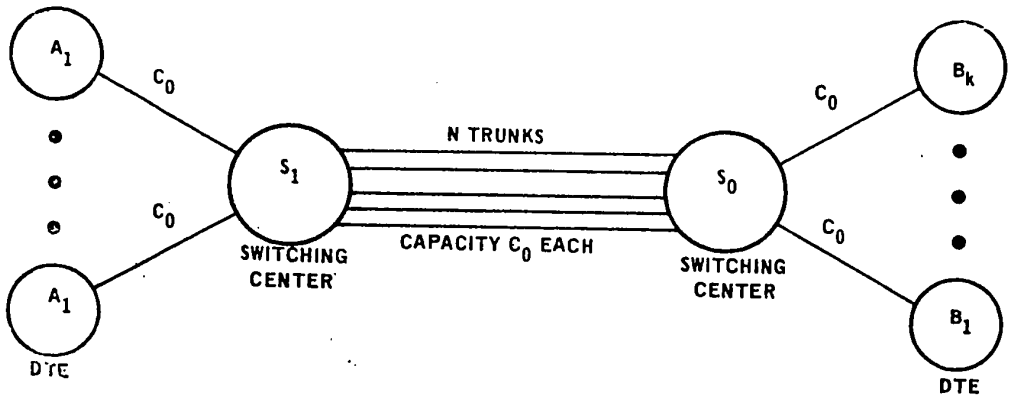

FIGURE 1
Circuit-Switching System

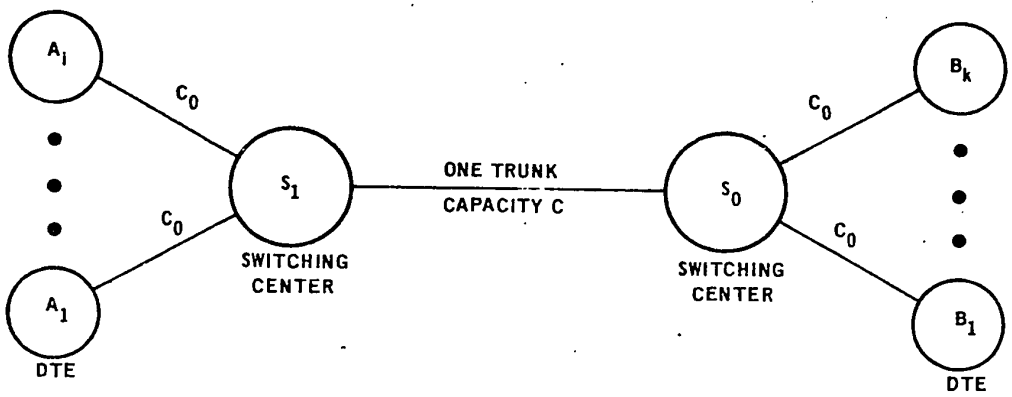

FIGURE 2
Message Switching

2.2 QUEUEING SYSTEM NOTATION

In describing queueing systems, the notation as proposed by Kendall° will be used. Hence a queueing system will be identified as $X/Y/Z$, where:

X = the input characteristic, i.e., the distribution of interarrival times of the customers.

Y = the distribution of the service time.

Z = the number of servers, an integer.

If

$X = M$, customers arrive at random (Poisson, Markov)

$X = D$, customers arrive at regular interval (deterministic)

$X = GI$, general independent input distribution.

$X = E_k$, Erlangian, a distribution which becomes M for $k = 1$ and D for $k = \infty$.

If

$Y = M$, the service time is negative exponentially distributed

$Y = D$, the service time is a constant

$Y = G$, no assumption about service time, general distribution

$Y = E_k$, Erlangian distribution.

The waiting time is the total time spent in the system, and it is equal to queueing time plus service time. Infinite queue means that an infinite waiting space is assumed for all calls that have to wait for their turn to be serviced.

2.3 ABBREVIATIONS AND SYMBOLS

b/s = bit per second = unit of transmission rate

c = transmission rate of a channel trunk

c_0 = transmission rate on the local loop

C_b = coefficient of variations

DTE = data terminal equipment

Δ_1 = difference in delay between $M/G/1$ solution and $M/M/1$ approximation

Δ_2 = difference in delay between $M/G/1$ solution and $M/D/1$ approximation

Δ_3 = difference in delay between $M/M/1$ and $M/D/1$ approximations

$= \Delta_1 + \Delta_2$

L = message length in bits

λ = average arrival rate of messages in messages per second

μ = average service time in seconds

μ_i = average service time in seconds in node S_i

N = number of trunks, servers (integer)

P_0 = Probability the queueing system with delay is empty

$$= \left[\sum_{n=0}^{N-1} \frac{(N\rho)^n}{n!} + \frac{(N\rho)^N}{N!(1-\rho)} \right]^{-1}$$

$P(\geq N)$ = Probability of encountering delay (Erlang C Formula)

$$= \frac{(N\rho)^N}{N!(1-\rho)} P_0$$

P_{0i} = Probability the queueing system with delay is empty at node S_i

$P_i(\geq N)$ = Probability of encountering delay at node S_i

Q = queueing delay = delay spent in systems before entering service

ρ = traffic intensity with $0 \leq \rho \leq 1$ for stable systems

ρ_i = traffic intensity at node S_i

S_0 = last node (switching center) before terminating DTE

S_i = node i $(i = 0,1,\cdots,3)$

T_0 = average time to transmit message on the local loop

T_D = fixed delay due to finite processing time of switch and propagation

T_{DN} = T_D/T_0 = normalized fixed delay

τ = connect time

τ_c = connect time for circuit switching system

τ_m = connect time for message switching system

$\tau(\cdots)$ = connect time for queueing system $(\cdots)$

W = waiting time = queueing time + service time

$\overline{W}$ = average waiting time

W_i = waiting time at node S_i

$\overline{W}_i$ = average waiting time at node S_i

$\overline{W}_{iN}$ = average waiting time at node S_i normalized by T_0

$X/Y/Z$ = notation of queueing system with

X = characteristics of input

Y = characteristics of service time

Z = number of servers (channels).

3. ONE-LINK SYSTEMS WITHOUT DELAY[1]

3.1 CIRCUIT SWITCHING

Assume that the system considered has the following properties:

a) it is a waiting system with infinite queue,

b) queueing discipline is first come, first served,

c) switching, signaling and propagation delays are neglected,

d) error free system, therefore no consideration for overhead for error correction and/or detection,

e) message arrivals at S_1 have a Poisson distribution with an average arrival rate of λ messages/second.

f) message lengths to be exponentially distributed with average message length equal to L bits.

g) transmission rate at subscriber's line c_0 b/s

h) number of interconnecting trunks N, each having a transmission rate of c_0 b/s.

Under the above assumptions, the model for the system is an *M/M/N* queueing system (i.e., random input, exponential service time, N servers) with first-come, first-served queueing discipline and the following characteristics:

a) mean arrival rate λ messages/second,

b) mean service time per message per channel $\frac{1}{\mu} = \frac{L}{c_0} = T_0$.

From queueing theory[3,5,6] we find that for this system, the average waiting time $\overline{W}$ is:

$$\overline{W} = \frac{P(\geq N)}{(1-\rho_1)\mu_1} + \frac{N}{\mu_1} \qquad (1)$$

where

$$\mu_1 = N\mu = Nc_0/L = N/T_0,$$

$$\rho_1 = \lambda/\mu_1 = \text{traffic intensity, } 0 \le \rho_1 \le 1,$$

$$P(\ge N) = \frac{(N\rho_1)^N}{N!(1-\rho_1)} \qquad P_0 = \text{Probability of a call encountering delay (Erlang } C \text{ formula).}$$

$$P_0 = \left[\sum_{n=0}^{N-1} \frac{(N\rho_1)^n}{n!} + \frac{(N\rho_1)^N}{N!(1-\rho_1)}\right]^{-1}.$$

From our definition of connect time, we have for this circuit switching system

$$\tau_c = \overline{W} - T_0 = \frac{P(\ge N)}{(1-\rho_1)\mu_1} = \frac{P(\ge N)}{(1-\rho_1)N/T_0}$$

Normalizing this connect time with respect to T_0 yields

$$\frac{\tau_c}{T_0} = \frac{P(\ge N)}{(1-\rho_1)N}. \qquad (2)$$

3.2 MESSAGE SWITCHING

To compare the circuit-switching system with the message-switching system, we consider the case where the total capacity of the interconnection trunks between S_1 and S_0 is c. We will have the same assumptions as in circuit switching with the modification that we have one trunk line between S_1 and S_0 with a transmission rate of c b/s.

The model becomes an $M/M/1$ system with

a) mean arrival rate λ messages/second,

b) mean service time per message = L/c,

c) number of servers $N = 1$.

Substituting these values into equation (1) we obtain for the average waiting time

$$\overline{W} = \frac{\frac{\rho_1}{1-\rho_1}\left[1+\frac{\rho_1}{1-\rho_1}\right]^{-1}}{(1-\rho_1)\mu_1} + \frac{1}{\mu_1}$$

$$= \frac{\rho_1}{(1-\rho_1)\mu_1} + \frac{1}{\mu_1}$$

with

$$\rho_1 = \lambda/\mu_1, \quad 0 \leq \rho_1 \leq 1,$$

$$\frac{1}{\mu_1} = L/c.$$

To compare this with the circuit-switching system, we assume the total trunk capacity in both cases to be the same, hence:

$$c = Nc_0$$

$$\frac{1}{\mu_1} = \frac{L}{c} = \frac{L}{Nc_0} = \frac{T_0}{N} .$$

Hence, by definitions (see Section 2), the normalized connect time for this message switched system is:

$$\frac{\tau_m}{T_0} = \frac{\rho_1}{(1-\rho_1)N} + \frac{1}{N} . \tag{3}$$

3.3 DISCUSSION

If we look at equation (2) for the connect time of the circuit-switching system then we can rewrite it as

$$\frac{\tau_c}{T_0} = \frac{1}{(1-\rho_1)N} \cdot \frac{1}{1+(1-\rho_1)\sum_{n=0}^{N-1}\frac{N!(N\rho_1)^{n-N}}{n!}} .$$

Since $0 \leq \rho_1 \leq 1$, clearly

$$(1-\rho_1) \sum_{n=0}^{N+1} \frac{N!(N\rho_1)^{n-N}}{n!} \geq 0.$$

Therefore,

$$\frac{\tau_c}{T_0} \leq \frac{1}{(1-\rho_1)N} .$$

For the message switching system, equation (3) can be rewritten as

$$\frac{\tau_c}{T_0} = \frac{\rho_1+(1-\rho_1)}{(1-\rho_1)N} = \frac{1}{(1-\rho_1)N} .$$

Hence

$$\tau_c \leq \tau_m .$$

We can therefore conclude that for the simple single-link system without delay the circuit-switching system has a smaller connect time.

4. THREE-LINK SYSTEM WITHOUT DELAY

In intercity traffic, a connection between terminal A and terminal B will most likely pass through more than two switching centers. Figure 3 shows, for example, a connection from DTE A to DTE B via A's serving local switching center S_3, toll centers S_2 and S_1 to B's local switching center S_0.

4.1 CIRCUIT SWITCHING

If we follow a circuit-switched connection between A and B, then we can see the following taking place:

a) At the originating office S_3, there is contention for a free trunk line to S_2, hence queueing takes place.

b) Once a free line is obtained, another queueing occurs at S_2. The trunk line between S_3 and S_2 is occupied although the message cannot be sent yet.

c) Upon seizure of a free line to S_1, another contention takes place at S_1. During this queueing, two trunk lines are being held busy.

d) If a free trunk to S_0 becomes available, and terminal B is not busy, then transmission of the message can begin.

The problem as stated above cannot be solved mathematically. To overcome this, we make the following additional assumptions

a) In all nodes, the number of arrivals will have a Poisson distribution. This can be justified if there are messages leaving the nodes and new arrivals coming to the node, all independent of one another (see Kleinrock's "Independence Assumptions"[3]).

b) All nodes have the same mean arrival rate and the same number of trunks. This assumption is not necessary, but it will make the calculation easier.

c) The holding time at S_1 is exponential with an average holding time per message L/c_0.

d) Node S_2 will now see a holding time consisting of the exponentially distributed holding time per message plus the queueing time at S_1. If we consider the traffic at S_2, it is quite impossible to calculate the distribution of holding times at S_2. We therefore make the assumption that this distribution is exponential with a mean holding time equal to the average waiting time $\bar{W}_1$ (i.e., queueing plus service time) at S_1. This assumption is not restrictive, especially for large numbers of trunks (see Palm[4], p.53-57; Cox[6], p.101-102; Syski[5]).

e) A similar assumption is made for node S_3: an exponential holding time with mean $\overline{W}_2$.

f) Then the average connect time at S_3 is equal to $(\overline{W}_3 - T_0)$.

Note that the assumptions d) and e) result in a model where the holding time progressively increases from node S_1 to S_2. Not all calls arriving at node S_3 will go over two additional nodes, but on the other hand, not all calls originating at S_1 terminate at S_0, over just one link. Port and Closs[1] justified this by substantiation with results of simulation studies.

Under these additional assumptions we can write out the following relations with the understanding that the index $i(i=1,2,3)$ refers to the node S_i; if the holding time per message is T_0, then

$$1/\mu = T_0 = L/c_0.$$

For the switching node S_1 we have

$$\mu_1 = N\mu = Nc_0/L = N/T_0,$$

$$\rho_1 = \lambda/\mu_1.$$

The normalized waiting time at node S_1 is, from equation (1),

$$\overline{W}_{1N} = \frac{\overline{W}_1}{T_0} = \frac{P_1(\geq N)}{(1-\rho_1)N} + 1 \tag{4}$$

with

$$P_1(\geq N) = \frac{(N\rho_1)^N}{N!(1-\rho_1)} P_{01}, \tag{5}$$

$$P_{01} = \left[\frac{(N\rho_1)^N}{N!(1-\rho_1)} + \sum_{n=0}^{N-1} \frac{(N\rho_1)^n}{n!}\right]^{-1}. \tag{6}$$

Under assumption d) the average waiting time $\overline{W}_1$ will become the service time for the N servers' system at S_2, therefore:

$$\mu_2 = N/\overline{W}_1.$$

$$\rho_2 = \frac{\lambda}{\mu_2} = \frac{\lambda}{\mu_1}\frac{\mu_1}{\mu_2} = \rho_1 \frac{N/T_0}{N/\overline{W}_1} = \rho_1 \frac{\overline{W}_1}{T_0} = \rho_1 \overline{W}_{1N}$$

Since $\overline{W}_{1N}$ can be calculated from equation (4), ρ_2 is, hence, known. Therefore

$$\overline{W}_2 = \frac{P_2(\geq N)}{(1-\rho_2)\mu_2} + \frac{N}{\mu_2} = \frac{P_2(\geq N)}{(1-\rho_2)N/\overline{W}_1} + \frac{N}{N/\overline{W}_1}$$

$$= \left[\frac{P_2(\geq N)}{(1-\rho_2)N} + 1\right]\overline{W}_1.$$

Normalized by T_0,

$$\overline{W}_{2N} = \frac{\overline{W}_2}{T_0} = \left[\frac{P_2(\geq N)}{(1-\rho_2)N} + 1\right]\overline{W}_{1N}.$$

$P_2(\geq N)$ and P_{02} are similar to (5) and (6) with ρ_2 substituted for ρ_1.

For node S_3 we can define

$$\mu_3 = N/\overline{W}_2$$

$$\rho_3 = \lambda/\mu_3 = \rho_1\overline{W}_{2N} \tag{7}$$

and hence

$$\overline{W}_{3N} = \left[\frac{P_3(\geq N)}{(1-\rho_3)N} + 1\right]\overline{W}_{2N} \tag{8}$$

with $P_3(\geq N)$ and P_{03} similarly defined as (5) and (6). The normalized connect time is

$$\frac{\tau_c}{T_0} = \overline{W}_{3N} - 1. \tag{9}$$

Equations (8) and (9) are too complex to be explicitly expressed in the original variables ρ_1, $1/\mu$, and N. However, the equations for the waiting times and the traffic intensities can be successively computed with the help of a computer. The results for the connect times are plotted in Figures 4 and 5 on a linear and logarithmic scale respectively.

It can be seen from equation (8) that $\overline{W}_{3N}$ has a pole for $\rho_3 = 1$. However, from equation (7) we have: $\rho_3 = \rho_1\overline{W}_{2N}$.

Hence, the pole occurs at $\rho_1 = 1/\overline{W}_{2N}$.

Since $\overline{W}_{2N}$ will be larger than 1, long before ρ_1 approaches 1, the pole of $\overline{W}_{3N}$ (or τ) will be $\rho_1 \ll 1$, but approaches 1 as the number of trunks N is increased.

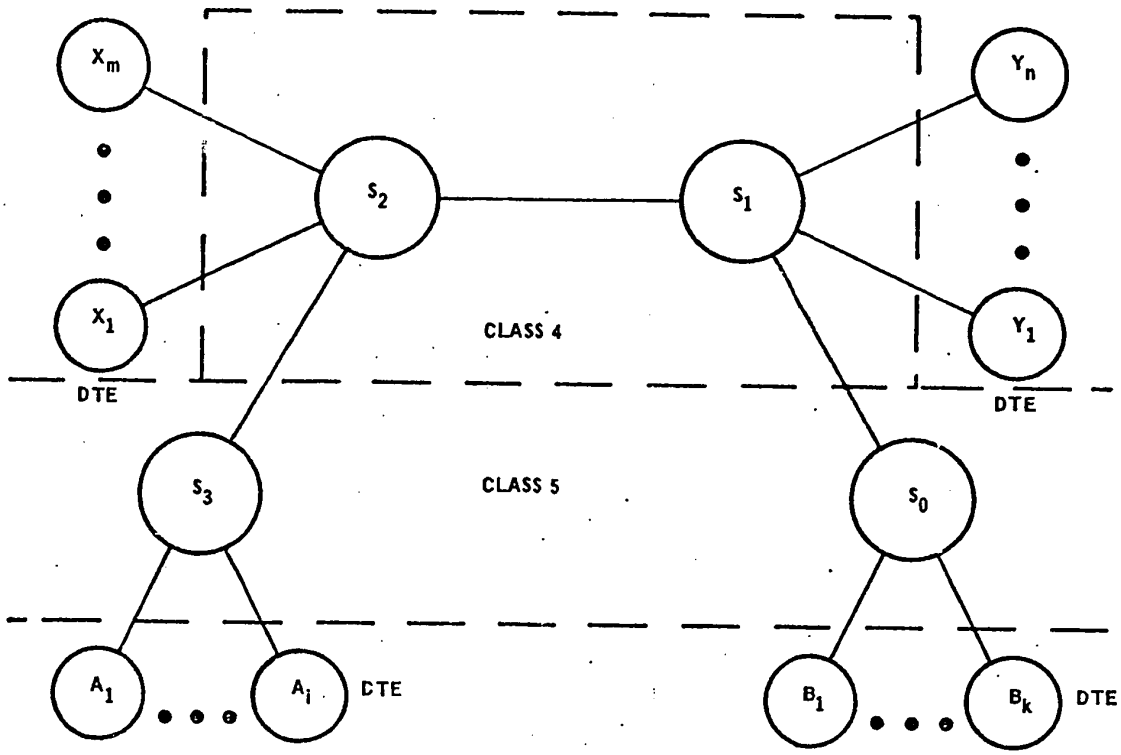

FIGURE 3
Three-Link Connection

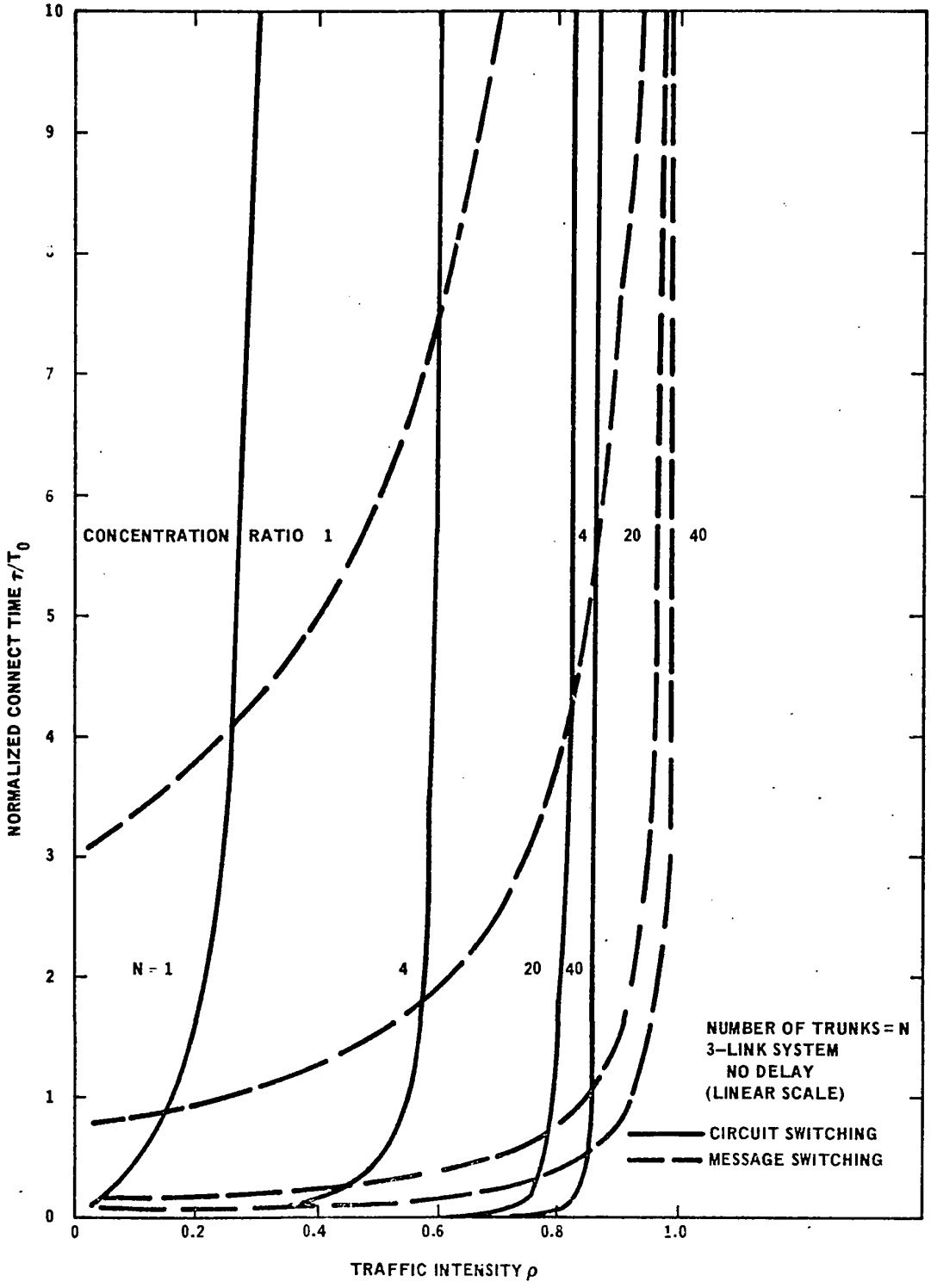

FIGURE 4
Connect Times

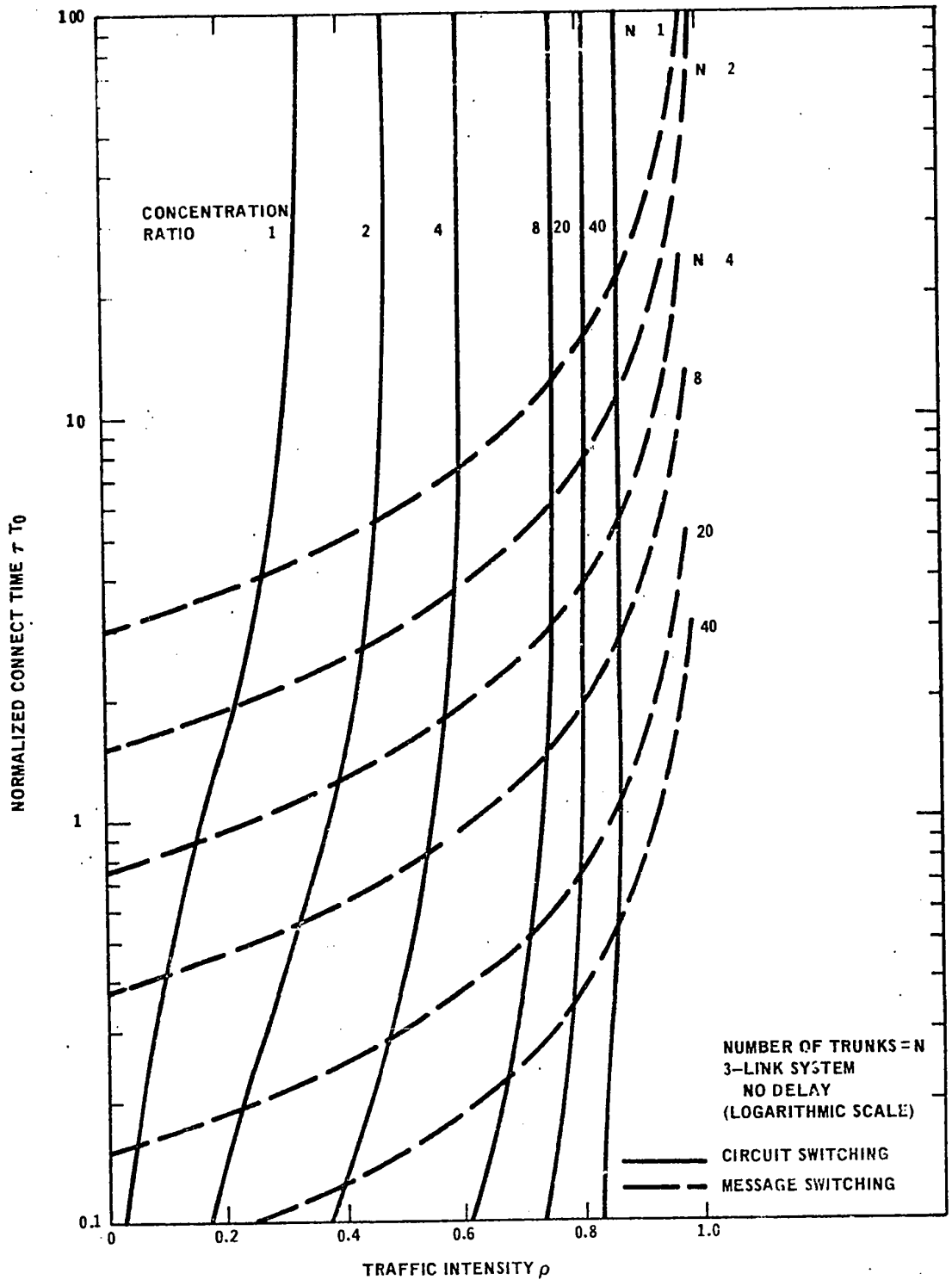

FIGURE 5
Connect Times

4.2 MESSAGE SWITCHING

If we now consider a message-switching system over three switching nodes similar to Figure 3, then the total transfer time to convey the message will be:

a) transmission time over the local loop from DTE A to S_3, plus

b) waiting times W_3, W_2, W_1, in nodes S_3, S_2, and S_1, plus

c) transmission time over the local loop from S_0 to DTE B.

Following the arguments in Section 2, we will neglect the transmission time from S_0 to DTE B, and for comparison with the connect time as calculated in Section 4, CIRCUIT SWITCHING, we consider as connect time

$$\tau_m = W_3 + W_2 + W_1 .$$

Taking the same assumptions, we will get, in this case, $W_3 = W_2 = W_1$ and from equation (3),

$$\frac{\tau_m}{T_0} = 3\left[\frac{\rho_1}{(1-\rho_1)N} + \frac{1}{N}\right] . \qquad (10)$$

This is also plotted in Figures 4 and 5 (broken lines).

4.3 DISCUSSIONS

As we can see from the graphs of Figure 4 or 5, for small incoming traffic intensity ρ_1, circuit switching is faster than message switching. However, as traffic builds up, the connect time in the message-switching system does not rise as fast as in the circuit-switching case. The pole of the latter is always for ρ_1 less than 1, while the message-switching system has a pole at $\rho_1 = 1$.

Clearly, if the connect time is already small, i.e., if its absolute value is small, then the much smaller connect time achieved by the circuit-switching system is not that important. However, its sensitivity to ρ as the traffic buids up is far greater than in the message-switching case. For a small number of trunks, the circuit-switching system can handle less traffic than the message-switching system, however, the difference becomes smaller with a large number of trunks (large N).

5. PRELIMINARY TO SYSTEMS WITH DELAY

5.1 ONE TRUNK, ONE LINK: *M/G/1*

Consider a one-server queueing system (i.e., one interconnecting trunk) where the message arrivals have a Poisson distribution with mean arrival rate λ, and the queueing discipline is first come, first served.

Its service time will consist of two parts:

a) a fixed delay T_D, which can be attributed to the finite switching time of the switching machine and to the finite propagation delay, and

b) a variable delay, which follows an exponential distribution (according to message length) with a mean service time $T_0 = L/c_0$.

Clearly this is an *M/G/1* queueing system. For this *M/G/1* system there exists an exact solution. However we would also like to know what error we make if we approximate this by an *M/M/1* or *M/D/1* system.

Exact Solution

For a general service-time distribution with finite first and second moment about the origins of service time, a coefficient of variation C_b is defined as follows:

$$C_b^2 = \frac{\sigma^2}{\overline{x}^2} = \frac{\overline{(x-\overline{x})^2}}{\overline{x}^2} = \frac{\overline{x^2}-\overline{x}^2}{\overline{x}^2}$$

where

x = service time,

σ^2 = variance of service time,

$\overline{x}, \overline{x^2}$ = first and second moment of service time.

In particular

C_b = 0 for a constant service-time distribution,

C_b = 1 for an exponential service-time distribution.

The average waiting time for this system is given by the Pollaczek-Khinchine formula: (see Cox[6] p.55)

$$\bar{W} = \frac{\rho\bar{x}(1+C_b^2)}{2(1-\rho)} + \bar{x} \tag{11}$$

where

$$\rho = \lambda\bar{x}.$$

For the problem as postulated, we have

$$\bar{x} = T_D + T_0,$$

$$C_b^2 = \frac{T_0^2}{(T_D+T_0)^2},$$

and

$$\bar{W} = \frac{\lambda(T_D+T_0)^2\left[1 + \frac{T_0^2}{(T_D+T_0)^2}\right]}{2[1-\lambda(T_D+T_0)]} + (T_D+T_0),$$

$$\frac{\bar{W}}{T_0} = \frac{\rho_0[1+(1+T_D/T_0)^2]}{2[1-\rho_0(1+T_D/T_0)]} + (1+T_D/T_0),$$

$\rho_0 = \lambda T_0$ = traffic intensity without delay.

Therefore, the connect time $\tau(M/G/1)$ for this $M/G/1$ system is given by

$$\frac{\tau(M/G/1)}{T_0} = \frac{\rho_0[1+(1+T_{DN})^2]}{2[1-\rho_0(1+T_{DN})]} + T_{DN} \tag{12}$$

where

$T_{DN} = T_D/T_0$ = normalized delay.

Approximation by $M/M/1$

Assume that we have approximated the $M/G/1$ system by an $M/M/1$ system, then we will get

$$\bar{x} = T_D + T_0,$$

$$C_b = 1,$$

and from equation (11),

$$\bar{W} = \frac{\lambda (T_D+T_0)^2 2}{2[1 - \lambda (T_D+T_0)]} + (T_D+T_0),$$

$$\frac{\bar{W}}{T_0} = \frac{\rho_0 (1+T_{DN})^2}{[1-\rho_0 (1+T_{DN})]} + (1+T_{DN}).$$

Hence, the normalized connect time for this $M/M/1$ approximation is

$$\frac{\tau(M/M/1)}{T_0} = \frac{\rho_0 (1+T_{DN})^2}{[1 - \rho_0 (1+T_{DN})]} + T_{DN} . \tag{13}$$

Consider the difference Δ_1 between the $M/M/1$ approximation and the exact solution. Then from (12) and (13), and normalizing by T_0,

$$\frac{\Delta_1}{T_1} = \frac{\tau(M/M/1) - \tau(M/G/1)}{T_0}$$

$$= \frac{\rho_0 (2T_{DN}-T_{DN}^2)}{2[1 - \rho_0 (1+T_{DN})]} . \tag{14}$$

Approximation by $M/D/1$

Another approach would be to approximate the $M/G/1$ by a constant service time system $M/D/1$, for which we then have

$$\bar{x} = T_D + T_0,$$

$$C_b = 0.$$

From equation (11)

$$\overline{W} = \frac{\lambda(T_D+T_0)^2 \times 1}{2[1-\lambda(T_D+T_0)]} + (T_D+T_0),$$

$$\frac{\overline{W}}{T_0} = \frac{\rho_0(1+T_{DN})^2}{2[1-\rho_0(1+T_{DN})]} + (1+T_{DN}).$$

The normalized connect time for this $M/D/1$ system is

$$\frac{\tau(M/D/1)}{T_0} = \frac{\rho_0(1+T_{DN})^2}{2[1-\rho_0(1+T_{DN})]} + T_{DN}. \tag{15}$$

From (12) and (15) we get for the difference Δ_2 between the exact solution and this $M/D/1$ approximation:

$$\frac{\Delta_2}{T_0} = \tau(M/D/1) = \frac{\rho_0}{2[1-\rho_0(1+T_{DN})]} . \tag{16}$$

Discussion

A general requirement for the system discussed to be able to reach an equilibrium condition is:

$$\rho_0(1+\tau) \leq 1.$$

This is quite clear, as any delay in switching and/or propagation adds to the time in which the system will be considered busy. The graphs for the $M/G/1$, $M/M/1$ and $M/D/1$ cases are plotted in Figure 6 for two values of ρ_0 equal to 0.1 and 0.5. The poles for $\rho = 0.1$ and 0.5 occur at $T_{DN} = 1$ and $T_{DN} = 9$ respectively. It is clear that the average connect time in the $M/G/1$ case with $0 \leq C_b \leq 1$ is bounded by the $M/M/1$ case on the upper end and by the $M/D/1$ case at the lower end. In fact, from (12), (13) and (15), we have

$$\tau(M/D/1) \leq \tau(M/G/1) \leq 2\tau(M/D/1).$$

If we now take the difference Δ_3 between connect times for the $M/M/1$ and $M/D/1$ system, then $\Delta_3 = \Delta_1+\Delta_2$.

Substituting (14) in (16) and normalizing by T_0,

$$\frac{\Delta_3}{T_0} = \frac{\Delta_1+\Delta_2}{T_0} = \frac{\rho_0(1+T_{ND})^2}{2[1-\rho_0(1+T_{DN})]} .$$

Then,

$$\frac{\Delta_1}{\Delta_3} = \frac{2T_{DN}+T_{DN}^2}{(1+T_{DN})^2} \quad \text{(independent of } \rho_0\text{)},$$

and also

$$\frac{\Delta_2}{\Delta_3} = \frac{1}{(1+\tau)^2} \quad \text{(independent of } \rho_0\text{)}.$$

For example, the delay of the *M/G/1* system will be half-way between those of the *M/M/1* and *M/D/1* system if

$$\Delta_1 = \Delta_2 = 1/2\ \Delta_3$$

and this will give a value of the normalized delay of

$$T_{DN} = \sqrt{2} - 1 = 0.414.$$

For

$$\frac{\Delta_1}{\Delta_3} \leq 0.25 , \quad T_{DN} \leq \frac{2\sqrt{3}-3}{3} = 0.155.$$

For

$$\frac{\Delta_2}{\Delta_3} \leq 0.25 , \quad T_{DN} \geq 1.$$

Hence for T_{DN} greater than 1, we can better approximate the *M/G/1* system by *M/D/1*, and for $T_{DN} < 0.155$, approximate it by *M/M/1*.

5.2 MULTIPLE TRUNKS, ONE LINK (*M/G/N*)

A theoretical solution to the *M/G/N* system was proposed by Pollaczek[7] and others (Syski[5], p. 363-377), but the solution is too complex. For large *N*, however, Pollaczek and also Palm[4] (p. 53, 57) came to the conclusion that the *M/G/N* system can be approximated by *M/M/N*. In our case where $0 \leq C_b \leq 1$ (recall $C_b = 0$ for constant service time, $C_b = 1$ for exponential service time) we can definitely bound the solution by the solutions for an *M/M/N* and *M/D/N* system. For an *M/D/N* system, the average queueing delay[5] is

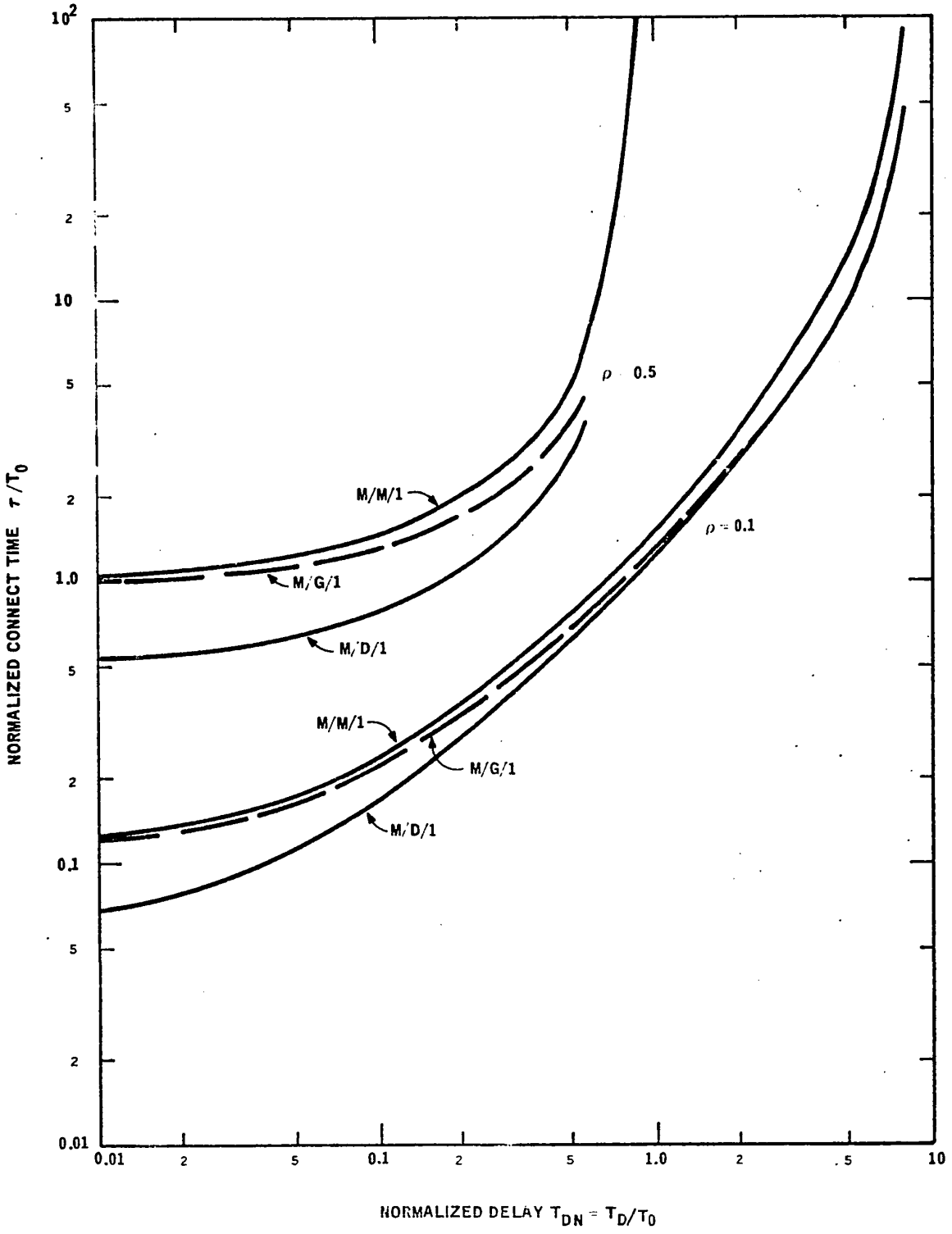

FIGURE 6

Connect Times for One Trunk, One Link Systems

$$\overline{Q} = \sum_{i=1}^{\infty} \exp(-iN\rho)\left[\sum_{x=iN}^{\infty} \frac{(iN\rho)^x}{x!} - \frac{1}{\rho}\sum_{x=iN+1}^{\infty} \frac{(iN\rho)^x}{x!}\right] \quad . \tag{17}$$

For an $M/M/N$ system, the solution can be derived from by equation (4) to be

$$\overline{Q} = \frac{(N\rho)^N}{N!(1-\rho)}\left[\sum_{n=0}^{N-1} \frac{(N\rho)^n}{n!} + \frac{(N\rho)^N}{N!(1-\rho)}\right]^{-1} \quad . \tag{18}$$

The queueing delay for constant service time is calculated on a computer using equation (17); for the exponential case the values are taken from a table.

In Figure 7, the ratio of the queueing delay for the two systems as given by equations (17) and (18) are plotted with the number of servers (trunks) N as a parameter and the traffic intensity ρ as a variable.

From this figure it can be seen that for large N and moderate traffic, the error of approximating an $M/D/N$ system by an $M/M/N$ system is not too severe. If however, the system considered is not an $M/D/N$, but an $M/G/N$ with a coefficient of variation C_b very close to 1, the coefficient of variation for an exponential service distribution, then the approximation of the service time distribution by a negative exponential will give a reasonable estimate of its performance.

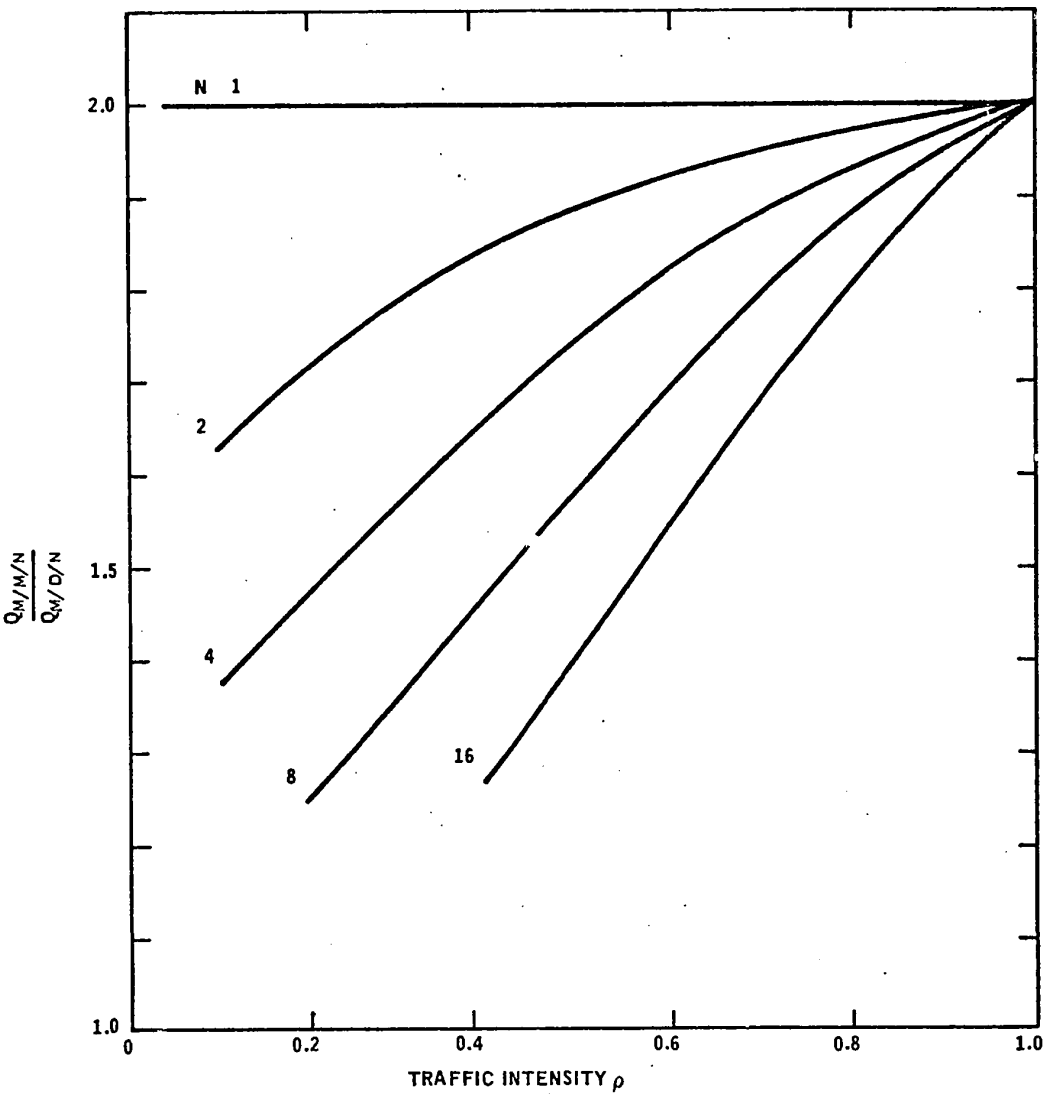

FIGURE 7

Ratio of the Queueing Delay for the two Systems

6. THREE LINK SYSTEM WITH DELAY

6.1 CIRCUIT SWITCHING

The analysis of the circuit switching system with delay will be based on a negative exponential approximation of the distribution of the service time. As shown in Section 5 for a reasonably large number of trunks and small values of C_b, this approximation will not introduce too much error. The analysis will be similar to that of Section 4.

Assume that the delay per switching link is T_D. Then the holding (or service) time for switching node S_1 (see Figure 3) is

$$\frac{1}{\mu_1} = \frac{T_D+T_0}{N} = \frac{T_0}{N}(1+T_{DN}) = \frac{1}{\mu_0}(1+T_{DN}) \tag{19}$$

with

$$\begin{aligned} T_{DN} &= T_D/T_0 = \text{normalized delay,} \\ \frac{1}{\mu_0} &= T_0/N = \text{service time without delay.} \end{aligned} \tag{20}$$

Then

$$\rho_1 = \frac{\lambda_1}{\mu_1} = \rho_0(1+T_{DN}) \tag{21}$$

with

$$\rho_0 = \frac{\lambda_1}{\mu_0} .$$

Here, as in Section 4, the index i (i=1,2,...) refers to the characteristics of node S_i. Hence, we get

$$\bar{W}_1 = \frac{P_1(\geq N)}{(1-\rho_1)\mu_1} + \frac{N}{\mu_1}$$

$$= \frac{P_1(\geq N)}{(1-\rho_1)N}(1+T_{DN})T_0 + (1+T_{DN})T_0 .$$

The normalized waiting time becomes

$$\overline{W}_{1N} = \frac{\overline{W}_1}{T_0} = \left[\frac{P_1(>N)}{(1-\rho_1)N} + 1\right] (1+T_{DN})$$

with $P_1(\geq N)$ and P_{01} as given by (5) and (6).

Proceeding as in section 4 with the second switching node S_2, we have the following equations

$$\frac{1}{\mu_2} = \frac{\overline{W}_1+T_D}{N} = \frac{\overline{W}_{1N}T_0+T_D}{N} = \frac{T_0}{N}(\overline{W}_{1N} + T_{DN}) = \frac{1}{\mu_0}(\overline{W}_{1N}+T_{DN})$$

$$\rho_2 = \frac{\lambda}{\mu_2} = \frac{\lambda}{\mu_2}(\overline{W}_{1N}+T_{DN}) = \rho_0(\overline{W}_{1N}+T_{DN}) .$$

Therefore

$$\overline{W}_2 = \frac{P_2(\geq N)}{(1-\rho_2)\mu_2} + \frac{N}{\mu_2}$$

$$= \frac{P_2(\geq N)}{(1-\rho_2)N}(\overline{W}_{1N}+T_{DN})T_0 + (\overline{W}_{1N}+T_{DN})T_0 .$$

Again normalizing by T_0

$$\overline{W}_{2N} = \frac{\overline{W}_2}{T_0} = \left[\frac{P_2(\geq N)}{(1-\rho_2)N} + 1\right](\overline{W}_{1N}+T_{DN}) .$$

For the switching node S_3 we get

$$\frac{1}{\mu_3} = \frac{\overline{W}_2+T_D}{N} = \frac{T_0}{N}(\overline{W}_{2N}+T_{DN}) = \frac{1}{\mu_0}(\overline{W}_{2N}+T_{DN})$$

$$\rho_3 = \frac{\lambda}{\mu_3} = \rho_0(\overline{W}_{2N}+T_{DN})$$

and

$$\overline{W}_{3N} = \frac{\overline{W}_3}{T_0} = \left[\frac{P_3(\geq N)}{(1-\rho_3)N} + 1\right](\overline{W}_{2N}+T_{DN}) .$$

$P_2(\geq N)$ and $P_3(\geq)$ are defined as in equation (5). The connect time normalized by T_0 is therefore

$$\frac{\tau_c}{T_0} = \overline{W}_{3N} - 1 \ .$$

In Figure 8, the normalized connect time τ_c/T_0 is computed as a function of $T_{DN} = (T_D/T_0)$ with ρ_0 as parameter for four different numbers of trunks, $N = 1,4,20,40$; in Figure 9 ρ_0 is taken as the variable and T_{DN} as the parameter.

6.2 MESSAGE SWITCHING

For the message-switching case, we will also use a negative exponential approximation for the service time distribution. Although in this case the calculation could be done without this assumption, the assumption is taken to make it comparable to the previous subsection. Therefore

$$\frac{1}{\mu_1} = (T_D+T_0/N) = T_0(T_{DN}+1/N) = (1+NT_{DN})T_0/N \qquad (22)$$

$$\rho_1 = \lambda/\mu_1 = \lambda(1+NT_{DN})T_0/N = \rho_0(1+NT_{DN})$$

$$\overline{W} = \frac{\rho_1}{(1-\rho_1)\mu_1} + \frac{1}{\mu_1} = \frac{\rho_0(1+NT_{DN})}{[1 - \rho_0(1+NT_{DN})]\ N/[T_0\ (1+NT_{DN})]} + \frac{1}{\mu_1}$$

$$\overline{W}_N = \overline{W}/T_0 = \frac{\rho_0[1+NT_{DN}]^2}{[1 - \rho_0(1+NT_{DN})]N} + \frac{1}{N} + T_{DN} \ .$$

Hence for the three-link system, the normalized connect time is

$$\frac{\tau_m}{T_0} = 3\overline{W}_N \ . \qquad (23)$$

The normalized connect times are calculated and plotted (broken lines) in Figure 8 with delay as a variable, and in Figure 9 with traffic intensity ρ as a variable.

6.3 DISCUSSIONS

Figures 8 and 9 show that for small N, message switching is able to carry more traffic than circuit switching. For large N circuit switching definitely has the advantage. However, in applying the graphs of Figures 8 and 9 the delay time T_D has to be interpreted correctly.

As stated in Section 2, a discrimination has to be made between

a) an average delay per message which is caused by the finite processing time of the switching machine, and

b) the propagation delay.

The first delay is very critical in message switching, the more so for a high concentration ratio N. This delay T_D is the delay to be used in equation (20) and will cause a very large waiting time for large N. Recall that the normalized delay T_{DN} is expressed in units of T_0, and therefore T_D itself becomes very large if compared with the service time, which is equal to T_0/N, in the high speed trunk of the message-switching systems. Neglecting propagation delay for the moment, it is obvious that, if the absolute delay T_D is the same for both message-and circuit-switching machines, the effect of this delay is more severe on message switching. Figure 9 (N=40) for example, shows this very clearly. It is therefore a definite requirement that the processing time for message switching be smaller than for circuit switching, if message switching is to be employed. If we assume that this processing delay is proportional to the number of instructions required to process a call, then existing packet-switchers as used in the ARPA-network[9] need about 500-600 instructions as compared to about 2000 instructions needed for electronic circuit switchers.

The effective processing delay in a message-switching machine can also be shortened by processing the next message, while simultaneously transmitting the current message on the high speed channel. Hence, as soon as this transmission is finished, the next message is immediately available for retransmission, thereby reducing T_D appreciably.

The second type of delay on the other hand does not significantly alter the connect time in message switching system. It does not affect the waiting time of the queueing model, only the time in the delivery of the message to the other subscriber. Hence, it becomes a term to be simply added to the expressions of the connect time in equation (21). On the other hand, in the case of circuit switching, this delay does affect the waiting time in the switching nodes. The delay T_D as in equation (19) has to include all delays caused by the finite propagation time during call set-up, thus also those due to signaling.

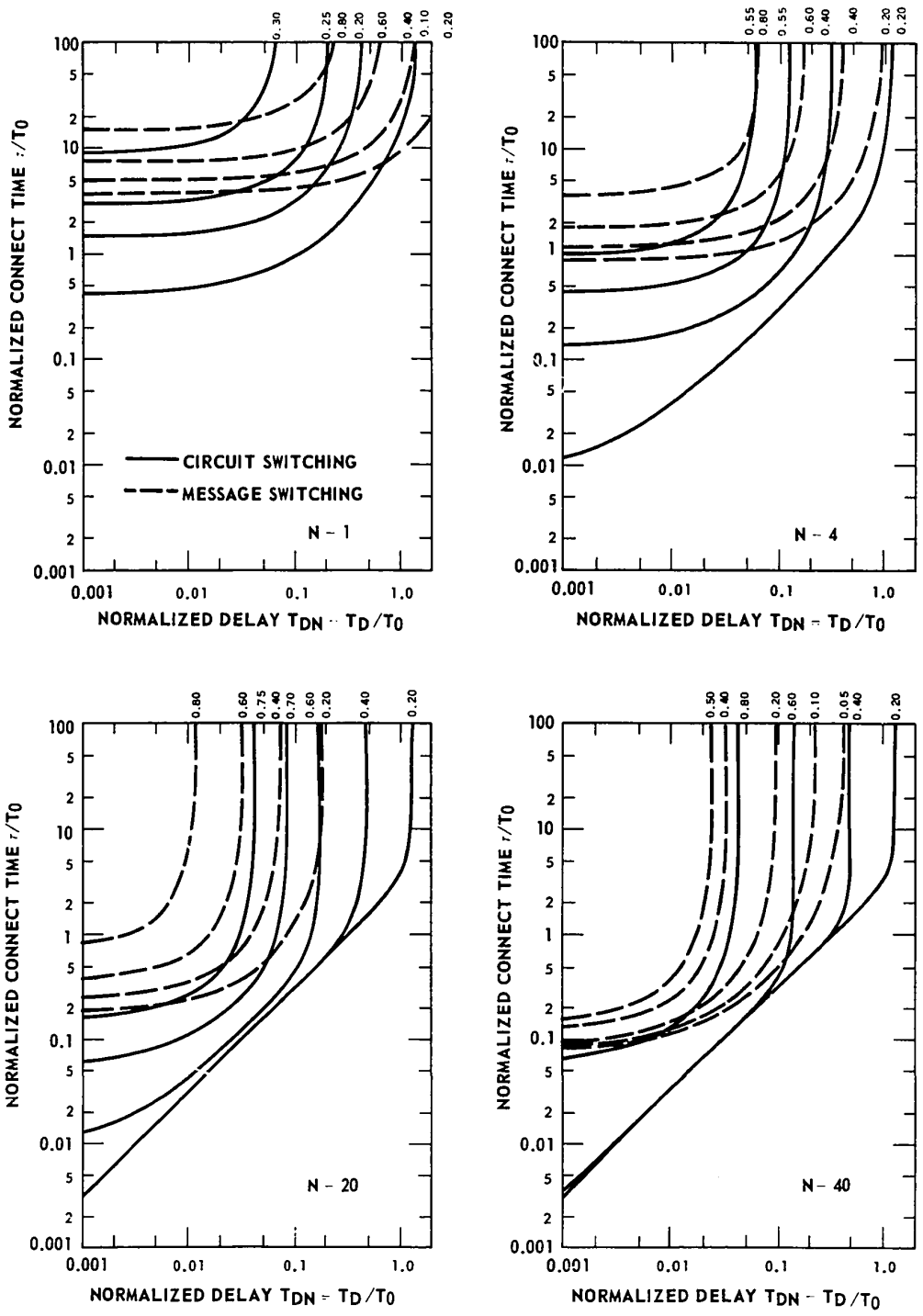

FIGURE 8

Connect Time vs. Delay (Parameter ρ)

If we consider Figure 9, the traffic intensity ρ is limited to value $\rho_{max} < 1$. The value $\rho = 1$ can be considered as the maximum number of messages (for the same value of average message length L) with no intermediate nodes and no delay. Hence

$$\rho_{max} = \frac{\text{max number messages in a switched system}}{\text{max number messages in a private line system with no intermediate nodes or delay}}.$$

ρ_{max} is therefore an indication of the utilization of the trunks.

6.4 EXAMPLES

Satellite Communications

EXAMPLE 1

Assume that

a) we have an average message of 1200 bits,

b) the local transmission rate = 1200 b/s, which gives $T_0 = 1$ s,

c) we have a trunk of total capacity 48 kb/s which we can split up in 40 trunks of 1.2 kb/s for circuit switching or let it be one trunk of 40 × 1.2 kb/s capacity for message switching,

d) switching delays of 5 ms per message per node,

e) propagation delay over satellite of 300 ms one-way.

The figures referred to below will be for N = 40.

Circuit Switching

Since the satellite link can be the first, the second or the last link, we will simplify the problem by dividing the delay due to propagation by three, and allocate the quotient to each link. Because of signaling requirement, the minimum delay caused by propagation is 2 × 300 ms = 600 ms. Therefore

$$T_D = \frac{600}{3} + 5 = 205 \text{ ms},$$

$$T_D/T_0 = 0.205.$$

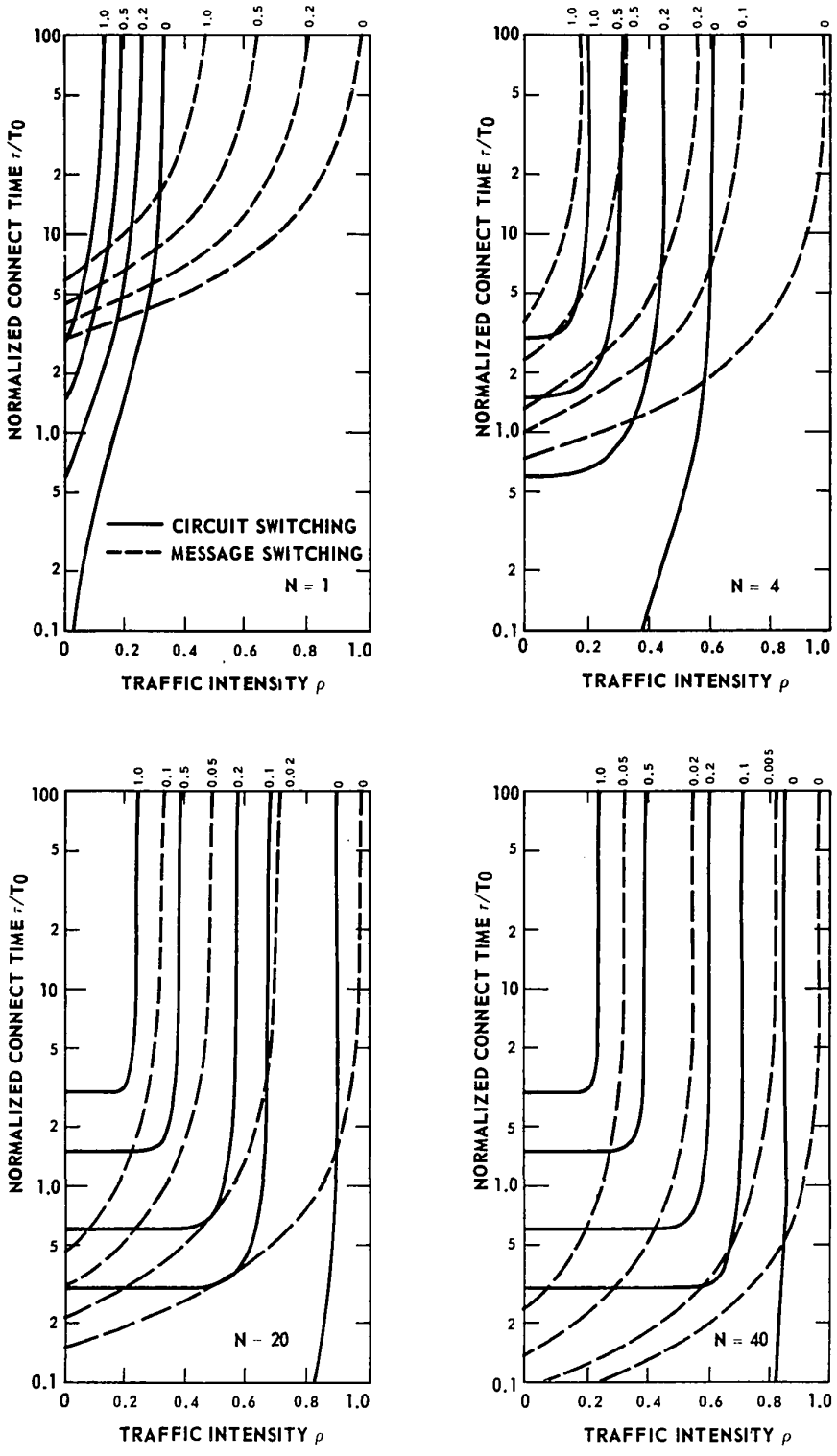

FIGURE 9

Connect Time vs. Traffic Intensity (Parameter τ)

From Figure 8, for $\rho = 0.50$, $\tau = 0.60 \times 1 = 0.60$ s. (see footnote)

From Figure 9, $\rho_{max} < 0.6$, and from (20) and (21), the maximum number of messages per second is

$$\lambda = N\rho_0/T = 24.$$

Messcge Switching In this case $T_D = 5$ ms, and $T_D/T_0 = 0.005$.

From Figure 8 for $\rho = 0.50$, $\tau = 0.23 \times 1 = 0.23$ s.

The total connect time = 0.23 + 0.3 = 0.53.

From Figure 9, for $T_D/T_0 = 0.005$, $\rho_{max} = 0.84$ and maximum number of messages per second is 33.

EXAMPLE 2

Same characteristics as in Example 1, except local transmission rate = 300 b/s which implies that $T_0 = 4$ s, and switching delay = 100 ms.

Circuit Switching

$$T_D = 200 + 100 = 300 \text{ ms},$$

$$T_D/T_0 = 0.3/4 = 0.075.$$

From Figure 8, for $\rho = 0.50$, $\tau = 0.22 \times 4 = 0.88$ s. (see footnote)

From Figure 9, for $T_D/T_0 = 0.075$, $\rho_{max} \simeq 0.75$.

Message Switching

$$T_D = 100 \text{ ms}$$

$$T_D/T_0 = 0.100/4 = 0.025$$

From Figure 8, for $\rho = 0.50$, $\tau = \infty$

Total connect time = ∞

From Figure 9, we can estimate ρ_{max} to be less than 0.50

Therefore the system yields infinite connect time for a traffic of 0.50.

In Example 1, message switching is comparable, in fact slightly better, to circuit switching if T_{DN} is small. However, as soon as the switching time is increased, then the advantage of circuit switching shows up as given by Example 2.

Strictly speaking the first data bit arrives 0.3 s later.

Conversational (Time Sharing) Mode

In most time sharing applications, the flow of messages is usually interrupted by "think" time. For a circuit switching system, the network cannot differentiate between think time and actual time when message is transmitted. Therefore, T_0 is much larger than the time actually needed.

For a message-switching system, no message is sent during the think time, and therefore the trunk is not unnecessarily occupied. If the think-to-message time ratio is 1:1, then ρ is already reduced by a factor of 2. In actual fact, this ratio will be much higher than two; a factor of 10 or more is probable. One can see how trunks can thus be better utilized by message switching and thus reducing the trunk transmission costs. (If the transmission cost becomes small compared to switching cost, a rating policy independent of distance is feasible).

REMARK. Because the information is sent by message switching (or for that matter by packet switching), this does not mean that the subscriber has to "redial" or "re-address" in order to send a second message to his correspondent, or that the latter cannot reply immediately to his query. Systems can be designed such that the end switching machines will automatically insert, after the initial setup, the addresses of the originating and destination subscribers. Thus as far as the subscribers are concerned, the system will still look as though it is "circuit switched."[8]

7. PACKET SWITCHING

In practice it may require less storage hardware per subscriber line to use packet switching than message switching. We could consider the following two kinds of packet switching.

1) The message is subdivided into packets. As soon as the message gets its turn to be served by the switching machine, the packets are sent sequentially, one immediately after the other. This system can be treated like message switching with an increase in processing time because of the header for each packet. Therefore, there is not much point considering it, since it does not provide any other advantage.

2) The message is subdivided into packets (possibly by the switching machine). As soon as the message enters service, one packet is processed; the rest is put back into queue to await its turn. Then another packet is processed, and the rest put back into queue again, and so on, until the whole message is processed. This system is called a round-robin service system (Kleinrock[5], Section 5.3). In its realization, this round-robin system may require less storage time and have faster so called "connect time".

For both systems, the calculation of average time spent in the system can be treated similarly as message switching, since the total amount of work is then the same. However, when one considers the distribution of waiting time in the round-robin case, this is quite different and will depend on the length of the message involved.

7.1 SYSTEMS WITHOUT DELAY

Let us consider the average waiting time first. Each message will be divided in packets, and each packet must be provided with a header. Hence, if the average message length is L, the header length L_H, and the packet length is L_P, then the message length is seemingly increased by

$$\frac{L}{L_P} L_H \ .$$

The traffic intensity is hence increased to

$$\rho' = \frac{\lambda(L+\Delta L)}{c_0} = \frac{\lambda L}{c_0}\left(1 + \frac{L_H}{L_P}\right) \ . \qquad (24)$$

If one considers the average waiting time at each node subjected to traffice ρ' as given by equation (24), then the average waiting time for the complete message in the packet switching mode is the same as in the message-switching mode with traffic ρ' (Kleinrock[3], Section 5.3). The average connect time per packet however, is much smaller. This will be

$$\frac{L_P}{L}\ \tau_m$$

where τ_m is given by equation (10).

The total connect time is therefore reduced by the same factor L_P/L. The probability of the total time spent in the system for the complete message is dependent on the message length. Messages with lengths shorter than the average length will spend less time, since the long messages are chopped into smaller packets, and between these packets a short packet message will slip through. Thus, it also follows that long messages will spend more time in the system since they will be more often interspersed by other messages.

7.2 SYSTEMS WITH DELAY

Again when one considers only average times, the approach as given in Section 6, MESSAGE SWITCHING, can be used. However, if T_D' is the delay to transmit a packet, then the total delay will be increased, namely by $T_D = L/L_P\ T_D'$. The reasoning in Section 7, SYSTEMS WITHOUT DELAY, can then be applied.

8. CONCLUSIONS

In conclusion we can summarize the results:

1. For a message-switching system it is imperative to keep the finite switching time low, such that it is much smaller than the average transmission time of a message over the trunk.

2. Propagation delay has no critical effect on message switching. In circuit switching this is not so; it will increase the holding time, and hence the connect time.

3. If information (message) flow is often interrupted by "think" time, then message switching will provide higher utilization of trunks.

4. Packet switching will decrease the connect time as compared to message switching, if the influence of finite switching time and the increase in traffic due to header are neglected. These two factors have an opposing effect. An optimum packet length must thus be found.

5. In the round-robin type packet-switching system, short messages will spend less time in the system.

6. For bulk messages, in which propagation delay can be neglected compared to message duration, and also where think time becomes negligible, circuit switching will prevail over message switching.

7. For a circuit-switching system over two or more links, the utilization factor of the trunks will increase with an increase in the number of trunks.

9. AREAS OF FURTHER STUDIES

To follow up this introductory study, the following areas could be further explored.

1. *SIMULATION* In the discussions in Sections 5 and 6, we have made certain assumptions. It would be very desirable to run a computer simulation to see how the results will be affected if some of the assumptions were dropped.

2. *PROBABILITY OF WAITING* So far only average times have been calculated. No attempt has been made to calculate the waiting time distribution.

3. *M/G/N SYSTEM* In Figure 6, the ratio of the average queueing delay for $C_b = 1$ to $C_b = 0$ was plotted against ρ. It would be interesting to see the plot for $0 < C_b < 1$.

4. *ADDRESSING AND CENTRALIZED CONTROL SIGNALING* In this report we have assumed that signaling is transmitted inband like the message, hence, the addressing information was transferred at the same speed as the message.

If, in a circuit switched environment we allocate N' trunks out of N for signaling, what will be the optimum number N', in order to maximize the flow of messages for a given waiting time? This question was answered for the case where the N' signaling trunks all have the same speed as the message. A more general question will be: given a total bit rate c b/s, how much capacity c'(b/s) do we allocate for signaling, and in how many signaling trunks N' should we subdivide c' to optimize the system? How does this optimized system compare to inband-signaling systems?

10. REFERENCES

1. E. Port and F. Closs, "Comparison of Switched Data Networks on the Basis of Waiting Times," *IBM Research Report* RZ 405 (#14721), Zurich, Switzerland, January 12, 1971.

2. Joint Working Party on New Data and Message Networks, "New Networks for Data Transmission: Current United Kingdom Position," *CCITT Joint Working Party New Networks for Data Transmission (GM/NRD) contribution* No. 2-E, Annex 1 (1968-1972).

3. L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*. McGraw-Hill, 1964.

4. C. Palm, "Research on Telephone Traffic Carried by Full Availability Groups," *Tele* (Ericson, A.B.) Vol. 1, 1957.

5. R. Syski, *Introduction to Congestion Theory in Telephone Systems*. Oliver and Boyd: London, 1960.

6. D.R. Cox and W.L. Smith, *Queues*, Methuen and Co. Ltd. London, 1961.

7. F. Pollaczek, "Uber das Warteproblem," *Math. Zeitschrift*, Vol. 38, p. 492-537, 1934.

8. D.G. kendall, "Stochastic Processes Occurring in the Theory of Queues," *Annals of Mathematical Statistics*, Vol. 24, pp. 338-354, 1953.

9. A. Brown, D. Gowan, D. Halliday, B. Han, and R. Prince, "Terminal Switch Interface in a Proposed Data Communications Network for Bursty Digital Data," *BNR internal report (draft)* February 10, 1972.

10. L.G. Roberts and B.D. Wessler, "Computer Network Development to Achieve Resource Sharing," *AFIPS Spring Joint Computer Conference Proceedings*, Vol. 36, pp. 543-549, May 5-7, 1970.

11. D.W. Davies, K.A. Bartlett, R.A. Scantlebury, and P.T. Wilkinson, "A Data Communications Network for Real-Time Computers," *IEEE Conference on Communications*, Philadelphia, pp. 728-733, 1968.

12. G.C. Hartley, "Data Communications Network and the Use of Store and Forward Methods," *Colloque International sur la Teleformatique*, Editions Chiron, 40 Rue de Seine, Paris 6e, France, Vol. I, pp. 348-357, March 24-28, 1969.

13. W.W. Chu, "Design Considerations of Statistical Multiplexors," *ACM Symposium on Problems in the Optimization of Data Communications Systems*, Pine Mountain, Georgia, October 13-16, 1969.

14. W.W. Chu, "Demultiplexing Consideration for Statistical Multiplexors," *ACM/IEEE 2nd Symposium on Problems in the Optimization of Data Communications Systems*, Palo Alto, California, October 20-22, 1971.