THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 7, September 1982
Printed in U.S.A.

A Variation on CSMA/CD That Yields Movable TDM Slots in Integrated Voice/Data Local Networks

By N. F. MAXEMCHUK

(Manuscript received January 6, 1982)

A variation on the carrier-sense, multiple-access/collision-detection (CSMA/CD) protocol for local-area, random-access broadcast networks is presented. The new protocol supports any mix of voice and data traffic, and has an upper bound on the delay of periodic traffic. In conventional CSMA/CD systems, statistical fluctuations in the transmission delay may cause disruptions in reconstructed voice waveforms. This does not occur with the new protocol. With this protocol, data sources use conventional CSMA/CD techniques to access the channel, while voice sources use a subset of these protocols, and appear to acquire a time-division multiplexed (TDM) slot. As in TDM systems, the entire system capacity can be used by the periodic sources. Unlike TDM systems, periodic slots are repositioned when other transmissions interfere with them, timing discrepancies between sources can be tolerated, and a centralized controller is not required to assign slots. Simulations are used to compare the delays in this system with those in conventional CSMA/CD systems. It is shown that at link utilizations of 0.9, voice sources can operate in a time-assignment, speech-interpolation mode. The effect of periodic traffic on conventional CSMA/CD protocols is also determined.

I. INTRODUCTION

There are basic differences between the requirements for data and voice transmission. There are also differences between the capabilities of local networks and the functions that should be performed on these networks, and general global networks. A protocol is described for integrating data and voice in a local-area, random-access broadcast network. It enables a single system, with similar interfaces for voice and data, to be used for both. The protocol satisfies the transmission requirements of both media, and takes advantage of the characteristics of local networks. This protocol can be applied to any system with periodic and aperiodic sources, as long as all of the periodic sources have the same transmission requirements.

In general, transmission capacity is less expensive in a local environment than in the global environment. This has resulted in significant differences in the types of networks implemented for data transmission in the two situations. In global networks, expensive access and switching techniques have been used to minimize the network capacity required to effect communications, whereas in local networks simpler switching techniques and less costly access nodes have been used. Several global networks, which use store-and-forward and packet-switching techniques to reduce the transmission requirements, are described in Chapters 1 and 2 of Ref. 1. A survey of techniques used in local networks, and an extensive bibliography is given in Ref. 2. Many of these local networks are configured as loops or random-access channels. Switching is simplified, but high-bandwidth communications channels are required.

The proximity of nodes in local networks leads to smaller propagation delays than those that occur in global networks. This has resulted in the development of more efficient random-access broadcast techniques for local networks[3,4] than have been developed for global networks (see Ref. 5 for a survey of random-access techniques and additional references).

Most proposals for integrated digital voice and data networks have applied to global networks. A survey of this work and an extensive bibliography can be found in Ref. 6. These proposals use variable-rate speech-coding techniques,[7,8] packet-switching techniques,[9,10] or complex strategies for combining circuit and packet switching.[11] These techniques are expensive to implement and should not be considered for local area communications. As in local data networks, integrated networks should take into account the reduced transmission costs and propagation delays in a local environment. In addition, only simple speech-coding techniques should be considered.

The requirements for digital voice transmission, both in terms of capacity and delay, are significantly different from those of data. The distribution of data traffic on local networks is typically bimodal,[12] comprising short interactive messages and long file transfers. The traffic from these sources arrives sporadically. If the message is divided into packets, a variance in the packet delay can be tolerated, providing the entire message delay is not excessive.

Uncompressed, digitized telephone calls require a large number of bits to be transferred, with much more stringent delay requirements than data sources have. Using a 32-kb/s speech coder, and transmitting only during active speech intervals, over 4-½ Mb of data must be transmitted during a 3-minute telephone call. The data cannot be accumulated over the entire call and then transmitted as a large file transfer between computers, because the participants interact. The maximum delay allowed in current telephone connections is on the order of a few hundred milliseconds. However, in a local network, the maximum delay must be significantly less than this, since the connection may also use an outside facility.

In a packet voice system, overhead bits must be transmitted in addition to the information from the digitized voice source. The more voice samples included in this packet, the higher the ratio of information to total bits in the packet, and the higher the transmission efficiency of the channel. However, the more voice samples included in a packet, the greater the delay between the time a sample is generated and the time it is delivered to the receiving telephone. As a compromise, packets consisting of several tens of milliseconds of speech are used in the systems described in this paper.

The variance of the delay in digital speech systems must also be constrained. The digital-to-analog converter at the receiver uses samples at a fixed rate. If a packet of samples is delayed to the extent that the previously transmitted samples are completely used up before it arrives, a discontinuity occurs in the speech. The probability of this occurring can be reduced by delaying the first packet of voice samples that arrives at the receiver and buffering future packets until they are needed. If the maximum delay is not constrained, this technique can reduce, but not eliminate, the problem. This delay adds to the overall delay between the speaker and the listener, and must be kept small.

In the global networks referenced earlier, packets of voice that do not arrive in time are lost. It is argued that if a small percentage of voice packets are discarded at random, the resultant distortion is tolerable. This same argument is used in proposals to integrate voice and data on local networks.[13,14] It may be valid to assume that voice packets are lost at random when aperiodic data sources generate most of the network requirements, or when the periodicity of speech sources is reduced by variable-rate speech-coding techniques. However, this assumption will probably not hold in a local network.

Based upon the measured data utilization of local networks[12] and the anticipated transmission requirements of digital voice, voice packets will most likely dominate integrated local networks. Since voice packets are generated periodically, if packets from voice sources collide, they are likely to continue colliding on successive transmissions. Therefore, successive delays from the same source will be correlated. Voice sources that do not contend with other sources may have a small average and variance of delay, while those that contend with a large number of voice sources may have a large average and variance of delay. If networks are designed based upon an acceptable average level of lost packets, and these packets are concentrated among a small number of connections during a small period of time, rather than being distributed randomly, the resultant service may not be acceptable.

Instead of reducing the periodicities to achieve fair packet losses, the periodicities can be used to eliminate lost packets entirely. This is accomplished by a variation on the carrier-sense, multiple-access/collision-detection (CSMA/CD) transmission protocol defined in Section II. This protocol transmits data by conventional CSMA/CD techniques, but uses only a subset of these techniques to transmit voice. The periodic sources do not detect collisions. In addition, periodic packets are tailored as described in Section III, the length of data packets is constrained, and periodic packets are given a higher retransmission priority. In Section IV, we show that this protocol limits the delay of voice packets to one data packet transmission time.

Periodic sources using this protocol operate as if a time-division multiplexed (TDM) channel has been assigned to each source. The difference between this channel and standard TDM channels is that it is not locked solidly into a time slot. The time slot may be shifted slightly back in time. When this occurs, the voice samples that arrive during the shift are transmitted in an expanded information area. An interesting characteristic of this system is that a periodic source can gain access to a system that does not appear to have the capacity to handle another periodic source. The system does not fail, but begins to operate as a completely utilized TDM system, with a slightly longer slot period. This phenomenon is explained in Section V. Another result of the time slot mobility is that timing discrepancies can exist between periodic sources without time slots being overwritten. This type of operation is described in Section VI.

In this system, if a periodic source and an aperiodic source are waiting to use a busy channel, the periodic source has a higher access priority and acquires the channel first. Typically, in queueing systems, when the delay of one class of users is reduced by raising its priority, the delay of the remaining users must increase. A simulator, which is described in Section VII, has been written to study the delays in this type of system. The model for the sources in the simulations is explained in Section VIII. The delays of periodic and aperiodic sources in this system are compared with the delays incurred when all sources use the same access protocol. These results are reported in Section IX. In Section X, alternative protocols are described.

II. TRANSMISSION PROTOCOL

In this protocol, all of the packets from an aperiodic source, and the first packet from a periodic source that is beginning to transmit, use a CSMA/CD protocol. Before transmitting, the source listens to the channel and refrains from transmitting if the channel is busy. While transmitting, the source also listens to the channel and stops transmitting if a collision with another source is detected. If the channel is busy, or if a collision occurs, the source tries again after the channel becomes idle. This protocol is not dependent upon the specific retry strategy these sources use to resolve contention. Any of the conventional strategies will work.

After the first transmission, a periodic source transmits all of the data it has accumulated whenever it acquires the channel. The source schedules its next transmission attempt at a fixed time T_p after the last successful transmission. In the first description of the transmission protocol, it will be assumed that T_p is the same for all sources. The effect of timing errors will be investigated in Section VI.

A periodic source listens before transmitting and defers transmission rights to terminals that are currently transmitting. However, this source does not listen while transmitting, and never terminates transmission prematurely. Instead, the packet structure for a periodic source is designed to allow aperiodic sources to detect a collision and terminate transmission before the periodic source begins transmitting useful data. When busy channels are encountered, the periodic source begins transmitting as soon as the channel becomes idle. This protocol, in conjunction with a constraint on the packet size from aperiodic sources, prevents packets from periodic sources from colliding.

As periodic traffic enters a system, the system becomes a TDM system. A periodic source acquires the channel and periodically uses a slot starting at this time until it is delayed by aperiodic traffic, or another periodic source starting to transmit. At this time, the periodic slot scheduled for the periodic source is shifted slightly. Additional data are transmitted in the first delayed slot to compensate for the shift.

III. PACKET STRUCTURE

The packets from aperiodic sources, Fig. 1a, consist of overhead bits and data bits. The overhead bits contain synchronization bits, source and destination addresses, packet-length counts, packet sequence numbers, error-control bits, and any other functions required by the communications protocols. These are variable-length packets. They are constrained to be less than or equal in length to the packets from the periodic sources.

Packets from periodic sources, Fig. 1b, consist of preempt bits, overhead bits, data bits, and overflow bits. During the preempt interval, the periodic source places a signal on the transmission media but

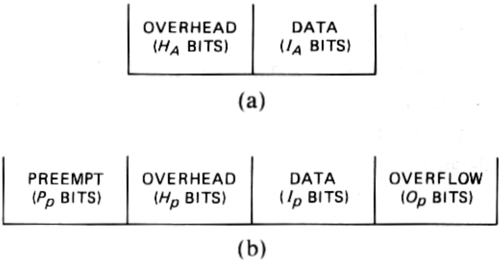


Fig. 1—(a) Packets from aperiodic sources. (b) Packets from periodic sources.

does not send information. This interval is long enough for an aperiodic source to detect a collision, stop transmitting, and have the effects of the transmission removed from the system before the periodic source begins transmitting useful data.

The length of the preempt interval, τ_p, is

$$\tau_p = 2\tau_t + \tau_{on} + \tau_I + \tau_{off},$$

where

τ_t is the maximum one-way propagation delay in the medium,

τ_{on} is the time required for a signal to be detected once it has propagated to a position on the channel,

τ_I is the time before the hardware reacts to a collision,

and

τ_{off} is time for a signal that is turned off to stop affecting a receiver, after the end of signal has propagated to the receiver.

Times τ_{on} and τ_{off} take into account electronics delays in the receivers and distortion in the received waveform caused by finite bandwidth channels which may not be correctly terminated. The preempt interval in a 3-Mb, 1-km simulated system is 38 bits long.

The overhead segment for periodic packets will be smaller than the overhead segment for aperiodic packets. An error-control strategy may not be used for periodic sources, since retransmitted packets will probably arrive too late to be used, and a greater error rate can be tolerated in sampled voice than in data. Packets will not arrive out of sequence; therefore, sequence numbers are not necessary. In addition, the number of bits of data in packets from these sources is deterministic, so that the length field is not required. In the simulations, $H_A =$ 100 bits and $H_p = 40$ bits.

When a periodic source acquires the channel, it transmits all the data it has accumulated. The source schedules its next transmission T_p seconds after a successful transmission. If the channel is not busy at this time, the samples it has collected fit in the data area. If the channel is busy, it waits before transmitting. The samples that arrive during this time are transmitted in the overflow area. Even when there are no overflow samples to be transmitted, the source transmits carrier during the overflow time. This guarantees that a periodic source takes no more time to transmit when it is delayed than when it acquires the channel immediately.

The size of the overflow area is determined by the maximum delay a periodic source can experience. This transmission protocol guarantees that the maximum delay for a periodic source will not exceed a packet transmission time. In the system simulated, the periodic sources generate 8000 4-bit samples per second, and T_p is 30 ms. A maximum of four samples arrive during a packet transmission interval; therefore, the overflow area is 16 bits.

The first packet from a periodic source may be shorter than successive packets since it does not include a preempt interval or an overhead interval. However, the same packet size will be maintained for this packet as for all other periodic packets. This guarantees that the scheduled interval between the next packet from this source and a packet from another periodic source is at least a periodic packet transmission time, X_p.

IV. CHANNEL CONTENTION

Delay is incurred in CSMA/CD networks when the channel is busy and when transmitting sources collide. Systems with periodic and aperiodic sources are being considered. Therefore, a periodic source can be delayed because

(*i*) the channel is busy transmitting an aperiodic packet,
(*ii*) a collision with an aperiodic source occurs,
(*iii*) the channel is busy transmitting a periodic packet, or
(*iv*) a collision with a periodic source occurs.

In this section, it is shown that only the first and third of these four mechanisms delay periodic sources, and that the maximum delay is less than X_p. It is assumed that every periodic source has the same T_p. The effects of timing inaccuracies are investigated in Section VI.

Whenever a periodic source and an aperiodic source collide, the aperiodic source detects the collision and stops transmitting before the periodic source begins transmitting useful data. Therefore, a periodic source is not delayed by a collision with an aperiodic source.

Whenever a periodic source and an aperiodic source are waiting for an idle channel, the periodic source obtains the channel. The periodic source begins transmitting as soon as the channel becomes idle. If an aperiodic source waits, it detects a busy channel and does not transmit. If an aperiodic source begins transmitting immediately, it detects a collision and stops transmitting. Therefore, a periodic source can only be delayed by an aperiodic source whose transmission is already in progress. This delay is at most the packet transmission time for an aperiodic source, X_a, which is constrained to be less than X_p.

Consider a sequence of k periodic sources scheduled to transmit at times

$$t_{1,1},\ t_{2,2},\ \cdots,\ t_{k,k}.$$

Let

$t_{i,j}$ be the time the transmission from source i is scheduled to appear at the location of source j ($t_{i,j} - t_{i,i}$ is the propagation delay from the ith to the jth source),

$t'_{i,j}$ be the time the transmission from source i actually appears at location j (this takes into account the channel acquisition delay for the ith source),

and,

$D_i = t'_{i,j} - t_{i,j}$ be the delay of the ith source. (Since the propagation time is independent of the delay, the delay between the scheduled and actual arrival is the same for each destination j, and the second subscript is not needed.)

The transmission from each periodic source lasts X_p, and each periodic source schedules its next transmission T_p after its last successful transmission. Therefore,

$$t_{i+1,i+1} - t_{i,i+1} \geq X_p.$$

If $D_i = 0$, the ith periodic source does not delay the $(i + 1)$th periodic source, and, as long as

$$D_i < X_p$$

these two sources will not collide.

Periodic sources cannot be delayed by periodic sources which have not been delayed. Therefore, the first periodic source to be delayed must be delayed by an aperiodic source. The delay incurred by the first periodic source is less than X_p. The effect of this delay may propagate and effect a sequence of periodic sources. In a general sequence of periodic sources, if

$$D_i < X_p,$$

then,

$$t'_{i,i+1} < t_{i+1,i+1},$$

and, the ith and $(i + 1)$th do not collide. The transmission time required by the ith source is X_p, even though it is delayed, and must transmit more samples. If

$$t'_{i,i+1} + X_p \leq t_{i+1,i+1},$$

then the $(i+1)$th is not delayed by the ith periodic source. This source may be delayed by an aperiodic source, and start a new sequence of delayed periodic sources, but the delay it incurs will be less than X_p. If

$$t'_{i,i+1} + X_p > t_{i+1,i+1},$$

the delay incurred by the $(i+1)$th source is

$$D_{i+1} = t'_{i,i+1} + X_p - t_{t+1,i+1}.$$

Since the $(i+1)$th source is waiting for the channel, this delay cannot be increased by an aperiodic source. The delay can be written as

$$D_{i+1} = D_i + t_{i,i+1} - t_{i+1,i+1} + X_p.$$

Since

$$t_{i,i+1} - t_{i+1,i+1} + X_p \leq 0,$$

$$D_{i+1} \leq D_i.$$

Therefore, the delay incurred by a sequence of periodic sources is a nonincreasing function, the maximum delay incurred by a periodic source is less than X_p, and periodic sources do not collide.

V. OVERLOAD TRAFFIC

Consider a system operating in a mode in which the channel capacity is nearly completely used by periodic sources. Assume that a time gap remains that is large enough for another source to begin transmitting, but not large enough to transmit an entire packet. Let another periodic source acquire the channel at this time. Is a failure mode created? Do delays build up until periodic sources start colliding? Is data lost? No. Instead, the system begins to operate without time gaps. The period between channel acquisitions increases, and some or all of the overflow bits in every packet are always used. Whenever a periodic source can acquire the channel, its ability to transmit is guaranteed.

Let a periodic source begin transmitting in a small time gap, such as that described in the hypothetical example. The periodic source it delays is delayed less than X_p. This source transmits the data accumulated during the delay in its overflow area and schedules its next transmission T_p seconds after it successfully acquires the channel. Successive periodic sources are delayed by an amount less than or equal to the delay incurred by the preceding source, as shown in the previous section.

The original interfering source becomes just another source in the sequence of interfering sources. It can be delayed by no more than the delay it originally caused, and can delay the source following it by no more than it did originally. Since the delay is a nonincreasing function, and it cannot go to zero for the over-utilized channel, it must stabilize at some positive time, ϵ, which is the same for all sources. The delay, ϵ, is equal to X_p, minus the sum of the idle channel times for a period T_p before the overflow source entered the channel. Using the terminology from the preceding section:

$$\epsilon = X_p - \sum_{t_{i,i+1} \epsilon T_p} [t_{i+1,i+1} - (t_{i,i+1} + X_p)].$$

When the stable situation occurs, each periodic source transmits a packet every $T_p + \epsilon$ seconds. It transmits the samples that have arrived in this period of time in the data and overflow area. At the end of each transmission, there is a periodic source that has been waiting ϵ seconds. This source acquires the channel before an aperiodic source or the first packet from another periodic source can. Until one of the sources ends its transmission, and channel capacity becomes available, the system operates as a TDM system with a slot period of $T_p + \epsilon$ seconds. No data are lost, and the slot delays do not grow indefinitely.

VI. TIMING CONSIDERATIONS

In a sampled data communication system, it is necessary for the transmitter and receiver to be frequency locked, so that samples are used at the same rate at which they are generated. In broadcast networks, this can be achieved by sending a clock signal outside of the data band or by using a modulation rule with a clock component. The former technique provides accurate timing, but requires that one unit be responsible for inserting the clock on the system. In the latter technique, there is no centralized control, every transmitting unit is identical, but timing discrepancies may exist between the transmitters, particularly when very little data are being transmitted.

Timing discrepancies result in the periodic terminals having different estimates of the interpacket interval T_p. Let the interpacket interval for the ith periodic terminal $T_{i,p}$ be within ϵ of T_p, so that

$$|T_{i,p} - T_p| \leq \epsilon.$$

Let the ith and $(i + 1)$th periodic terminals transmit at

$$t_{i,i+1} = t,$$

and,

$$t_{i+1,i+1} = t + X_p,$$

so that there is no separation of the packets at the $(i + 1)$th source. The next packets from these terminals are scheduled at

$$t_{i,i+1} = t + T_{i,p},$$

and,

$$t_{i+1,i+1} = t + X_p + T_{i+1,p}.$$

These two times may be separated by as little as $X_p - 2\epsilon$. If the first packet in this sequence is delayed by a packet from an aperiodic source, it may be delayed until

$$t'_{i,i+1} < t + T_{i,p} + X_a.$$

With the constraint $X_a \leq X_p$, it is possible that both periodic sources will be waiting for the channel and will collide. This situation cannot be resolved by the protocol. It can be prevented by constraining the length of aperiodic packets to

$$X_a \leq X_p - 2\epsilon.$$

With this constraint, it can be shown, as in Section IV, that sequences of periodic sources do not collide, and that the delay of a periodic source does not exceed X_p.

This length constraint cannot be used to solve all of the problems that arise because of timing inconsistencies. The first packet from a periodic source must have length X_p to reserve an entire slot, but must enter the system like a periodic packet. This can be resolved by allowing periodic sources to recognize and preempt the first packet from a periodic source that begins transmitting within 2ϵ of the periodic source's scheduled transmission. Now, the maximum delay caused by the first packet from a periodic source is not greater than that caused by a packet from an aperiodic source.

VII. SIMULATION

A general-purpose simulator has been written to determine the characteristics of broadcast networks. In this simulator, the protocols used by sources, the traffic generated by a source, the physical position of a source on a channel, the parameters that define a transmission channel, and the interconnection of channels can be varied.

In the simulator, time jumps between the occurrence of significant events, rather than continuously moving forward by small increments. Functions such as wait for idle channel and listen while transmitting are implemented by giving the channel the characteristics of an active device. The channel notifies a source when the state of the channel, at the location of the source, has changed. Actually, the channel in this type of system is passive and sources that perform these functions examine the channel continuously. However, the interval of time in which the state of the channel can change and a significant event can occur is small compared to the average time between significant events. A simulator that models the continuous operation of the sources would have to examine the channel at small intervals, and would be much less efficient, in terms of processing, than the technique selected.

In this simulator, the channel maintains lists of sources that are transmitting, listening while transmitting, and listening for an idle channel. When the state of the channel changes, channel maintenance routines notify the terminals in the lists at the appropriate times. These routines take into account the propagation delays between terminals, and the detection time of terminals.

Queues of messages and packets waiting to be transmitted from the sources are maintained without creating physical queues. Instead, the clocks associated with packet and message generation are allowed to run independent of the simulation clock. Each source keeps track of the time the packet and message it is transmitting were generated. When the transmission is complete, this time is updated to determine the next packet or message arrival at this source. If this time is less than the simulation time, this object has been waiting in a queue, and can be transmitted immediately. If it is greater than the simulation time, the queue is empty, and the message or packet is scheduled to arrive in the future.

The periodic terminals approximate voice terminals that only transmit during active speech intervals. For the parameters selected in the simulations, the speech interval durations are more than three orders of magnitude greater than the packet transmission time, and more than six orders of magnitude greater than certain of the channel-related events. The simulator tracks the fine structure of channel-related events. Because of the different time scales, the simulations could not be conducted for a long enough time to guarantee that all combinations of active and inactive intervals occur with the proper probabilities. To compensate for this, a different random-number generator is used to generate active and silent intervals than is used to generate the other random events in the simulation. This allows different protocols for periodic sources to be compared when the same active and silent intervals occur.

The systems simulated have the following characteristics:

(*i*) Sources are uniformly distributed along a single cable.

(*ii*) Channel length is 1 km.

(*iii*) Transmission rate is 3 Mb/second.

(*iv*) Time to detect the presence and removal of carrier, τ_{on} and τ_{off}, is 1 μs.

(*v*) Time to detect interference, τ_I, is 4 μs.

(*vi*) Time the channel must remain idle between transmissions is one-third of a microsecond.

These parameters do not correspond to a particular hardware realization of a system, but are indicative of what might be expected.

VIII. SOURCE MODELS

Aperiodic and periodic terminals are simulated. They correspond to a number of data sources sharing a store and forward node, and a two-way voice conversation. Some of the parameters associated with these terminals are selected to study specific characteristics of the system or to facilitate programming, rather than to accurately model the source. The results of the simulations provide an indication of how the system operates and not precisely how many terminals of a specific type can be supported.

The model for the sources in this system is shown in Fig. 2. In the remainder of this paper, the packets from the various sources are referred to as follows:

Type A—all of the packets from aperiodic sources.

Type F—the first packets in active intervals from periodic sources.

Type P—all other packets from periodic sources.

Type X—a composite of all of the packets from periodic sources.

Type T—a composite of all of the packets from all sources.

The protocols for periodic sources are referred to as follows:

Π_A—Periodic sources use the protocol proposed in this paper.

Π_L—Periodic sources use the same protocol as data traffic and transmit enough overhead bits to have the same packet size as Π_A.

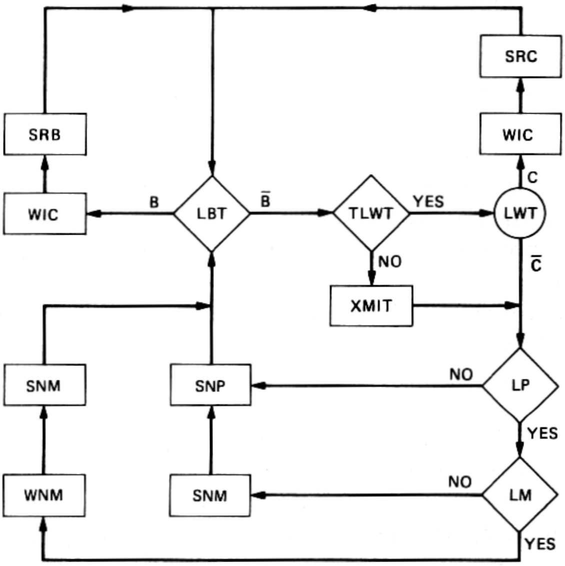
Π_S—Periodic sources use the same protocol as data traffic, but transmit fewer overhead bits than Π_L.

Π_X—Either Π_L or Π_S. This symbol is used when describing the intersection set of these two protocols.

The doublet (x, y) refers to packets of type x, when the periodic sources use protocol y.

In the model, all of the sources test the state of the transmission medium before transmitting (LBT). If the channel is currently being used (B), they wait until it is idle (WIC), then schedule a transmission retry according to a source-specific strategy (SRB). Transmission priorities can be established by using different strategies for different sources. For instance, if one type of source can retry sooner than another, these sources are more likely to find the channel idle and will have a smaller average delay. For this set of simulations, all of the sources begin transmitting as soon as the channel becomes idle. If the channel is not heavily used, there is seldom more than one source waiting, and this will lead to the smallest possible delay. If more than one source is waiting, they collide and the collision retry strategy determines which source obtains the channel.

Transmitting sources may or may not listen while transmitting. In this set of simulations, sources listen while transmitting type A, (F, Π_A), and (X, Π_X) packets. When a collision is detected, a source-



◇ TEST–TAKES NO TIME.

▭ EVENT WHICH MAY TAKE NONZERO TIME.

○ CONDUCTS A TEST WHILE AN EVENT IS IN PROGRESS.

LBT – LISTEN TO THE CHANNEL BEFORE TRANSMITTING.
LWT – LISTEN TO THE CHANNEL WHILE TRANSMITTING.
XMIT – TRANSMIT.
WIC – WAIT FOR IDLE CHANNEL.
SRB – SCHEDULE A RETRY AFTER A BUSY CHANNEL.
SRC – SCHEDULE A RETRY AFTER A COLLISION.
SNP – SCHEDULE TRANSMISSION OF NEXT PACKET.
SNM – SCHEDULE TRANSMISSION OF NEXT MESSAGE.
B – CHANNEL BUSY.
$\overline{B}$ – CHANNEL NOT BUSY.
C – COLLISION DURING TRANSMISSION.
$\overline{C}$ – NO COLLISION DURING TRANSMISSION.
LP – TEST IF THE TRANSMITTED PACKET WAS THE LAST PACKET IN THE MESSAGE.
LM – TEST IF THE TRANSMITTED MESSAGE WAS THE LAST MESSAGE IN THE QUEUE.
WNM – WAIT FOR NEXT MESSAGE.
TLWT– TEST IF THIS TERMINAL IS TO LISTEN WHILE TALKING.

Fig. 2—Model of the transmission strategies used by the sources in the simulations.

dependent retry strategy (SRC) is initiated. The sources do not listen while transmitting type (P, Π_A) packets.

The algorithm, SRC, provides another opportunity to assign transmission priorities. Sources generating type A and (X, Π_X) packets distribute their retrys uniformly over an interval $2^i\Delta\tau$, where i is the number of collisions this packet has experienced, and $\Delta\tau$ is two round-trip transmissions delays. The value of i is limited to ten, so the retry interval varies from two to 1024 round-trip delays. Sources generating type (F, Π_A) packets distribute their retrys uniformly over an interval Δ_T. Under heavy-load conditions, these souces retry sooner than those generating type A packets and have a smaller average delay. The periodic sources model voice terminals that only transmit during active speech intervals. Type (F, Π_A) packets are given a higher priority because they must acquire the channel in less than 50 ms more than 98 percent of the time.[15,16] It is shown in Section IX that this requirement is met at channel utilizations of 0.9.

In this model, after a packet is successfully transmitted the source determines if this is the last packet in a message. If it is not, the next packet transmission is scheduled. If it is, the next message is scheduled. For the periodic sources, a message corresponds to an active interval during which packets are generated. After the last packet in an active interval is transmitted, there is never a message waiting. Instead, a silent interval is generated, and lasts until the next message arrives. Type (P, Π_X) packets are generated every T_p seconds after the first packet successfully acquires the channel. These packets are queued until they can be transmitted. Type (P, Π_A) packets are generated T_p seconds after the last successful channel acquisition.

The aperiodic sources generate variable-length messages. The messages arrive independently and are queued until they can be transmitted. The message length determines the number of packets in the message and the length of the last packet. The source transmits all of the packets in a message before determining if another message is waiting. The packet scheduling algorithm (SNP) waits before transmitting successive packets from this source. This enables other sources to acquire the channel.

The aperiodic sources in these simulations have the following characteristics:

(*i*) The message interarrival process is a negative exponential distribution with a mean of 33.47 ms.

(*ii*) The packet size is 1050 bits, 90 overhead bits, and 960 data bits.

(*iii*) The message length is deterministic—it is 960 bits. (Each message generates a single maximum-size packet.)

(*iv*) The minimum time between successive packets from the same source is four round-trip transmission delays.

These characteristics do not correspond to any known data source. They are selected to allow comparisons between systems with different mixes of periodic and aperiodic sources. The average number of bits transmitted by one of these sources is equal to the average number of bits transmitted by a periodic source. This allows the network utilization to remain constant, while the mixture of periodic and aperiodic sources is varied. From this, the effect of periodic rather than aperiodic requirements on a broadcast network is determined. Using these characteristics, type A packets are almost the same size as type (X, Π_A) packets. This causes the largest possible delay to be incurred by the periodic sources.

The periodic sources have the following characteristics:

(*i*) For all of the sources, T_p is set to 30 ms. This results in 960 data bits per packet.

(*ii*) The duration of silent and active intervals are exponentially distributed with means of 0.185 and 1.31 seconds, respectively.

(*iii*) For type (X, Π_A) packets, $H_p = 40$ bits, $O_p = 16$ bits, and $P_p = 38$ bits. This results in 1054-bit packets.

(*iv*) For type (X, Π_L) packets, $H_p = 94$ bits. This retains the same link utilization as aperiodic sources and periodic sources using the new protocol.

(*v*) For protocol Π_s, $H_p = 40$ bits. This shows the effect of increasing the link utilization in the new protocol.

The mean values of the distributions are selected so that the total active time, the average active interval, and the number of active intervals correspond to those in a two-way telephone conversation.[17] This model assumes that the two speakers in a conversation do not talk simultaneously. It ignores the 0.07 probability of double talk. This model is easier to implement than one which allows double talk and should provide a reasonable indication of the system delays.

IX. SIMULATION RESULTS

Simulations have been conducted for protocols Π_A, Π_s, and Π_L. The fraction of the requirements generated by periodic sources was set to 0, 10, 25, 50, 75, 90, and 100 percent of the total requirements and the nominal offered utilization, S, to 0.7, 0.8, and 0.9. For protocols Π_A and Π_L, the actual values of S are very close to the nominal values. For protocol Π_s, the utilization generated by aperiodic sources and the number of periodic packets is the same as for the other protocols. However, the overhead associated with periodic packets is reduced. Therefore, the actual value of S is less than the nominal value.

Statistics were taken for 10 seconds of elapsed time after a 2-second initialization period. As the utilization varied from 0.7 to 0.9, the total number of packets in the simulations varied from about 20,000 to 25,500. The time between successive starts of active intervals for the periodic traffic is on the order of a second or two. It is unlikely that all combinations of active and inactive periodic terminals are witnessed in a 10-second interval. However, the same active intervals occurred for each of the protocols, and the data show definite trends. The average and standard deviation of the channel access delay are plotted as a function of the percentage of the load generated by periodic sources in Figs. 3 and 4.

For Π_L, the periodic and aperiodic sources use the same transmission protocol and have the same packet lengths. For this protocol, the network utilization remains constant as the load shifts from aperiodic to periodic sources. All of the curves (T, Π_L) display the same characteristic. For a small fraction of periodic load, the delay measures increase. As the fraction of the load generated by periodic sources becomes large, the delay measures decrease.

This behavior can be accounted for by characteristics of periodic sources. As the periodic sources change between the active and inactive mode, the short-term utilization changes. The delay versus utilization curve for this type of system is concave upward so that the delay averaged over a number of utilizations is greater than the delay associated with the average utilization. This accounts for the initial increase in the curves. The periodic sources transmit packets at a multiple of T_p after the first successful transmission. This separates sources so that a large number of sources do not collide. Since it takes longer to resolve contention between larger numbers of sources, the average delay decreases as the number of sources that may be separated increases. This accounts for the decrease in the delay measures when the fraction of the load generated by periodic sources exceeds ten percent.

Curves (T, Π_A) show the delay characteristics for all of the packets in a system that uses the new protocol. The total utilization is the same as for Π_L. The delays for (T, Π_A) are less than those for (T, Π_L). Protocol Π_A guarantees that periodic sources do not collide with each other and that they successfully acquire the channel when they collide with aperiodic sources. Since there are fewer collisions, and fewer of the collisions that occur result in unsuccessful transmissions, less of the channel capacity is wasted. The decrease in delay is a result of the reduced-channel contention.

Comparing Π_A and Π_L demonstrates the favorable characteristics of the new protocol but not the unfavorable characteristics. Protocol Π_A requires additional bits to be transmitted in the preempt and overflow areas of the message. To show the effect of these bits, Π_s is examined. This protocol is the same as Π_L, except the overhead in periodic packets is reduced. As the fraction of the load generated by periodic sources increases, the average channel utilization decreases. Curves (T, Π_s) are below (T, Π_L), but are still above (T, Π_A). The latter result is due to the ratio of data bits to packet length in the system simulated. The preempt header is dependent upon the propagation delay in the system. As the channel becomes longer, or the transmission rate

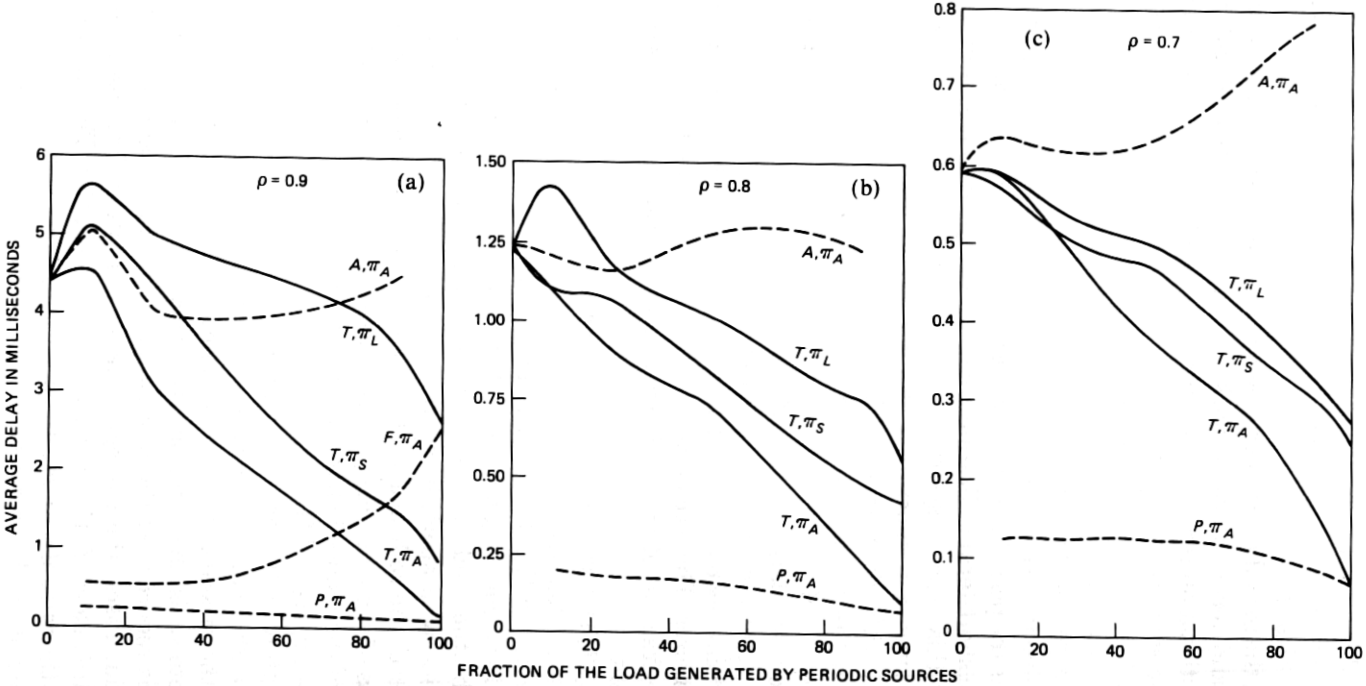


Fig. 3—Average delay versus the fraction of the load generated by periodic sources for several types of sources with protocols at the following utilizations: (a) $\rho = 0.9$; (b) $\rho = 0.8$; (c) $\rho = 0.7$.

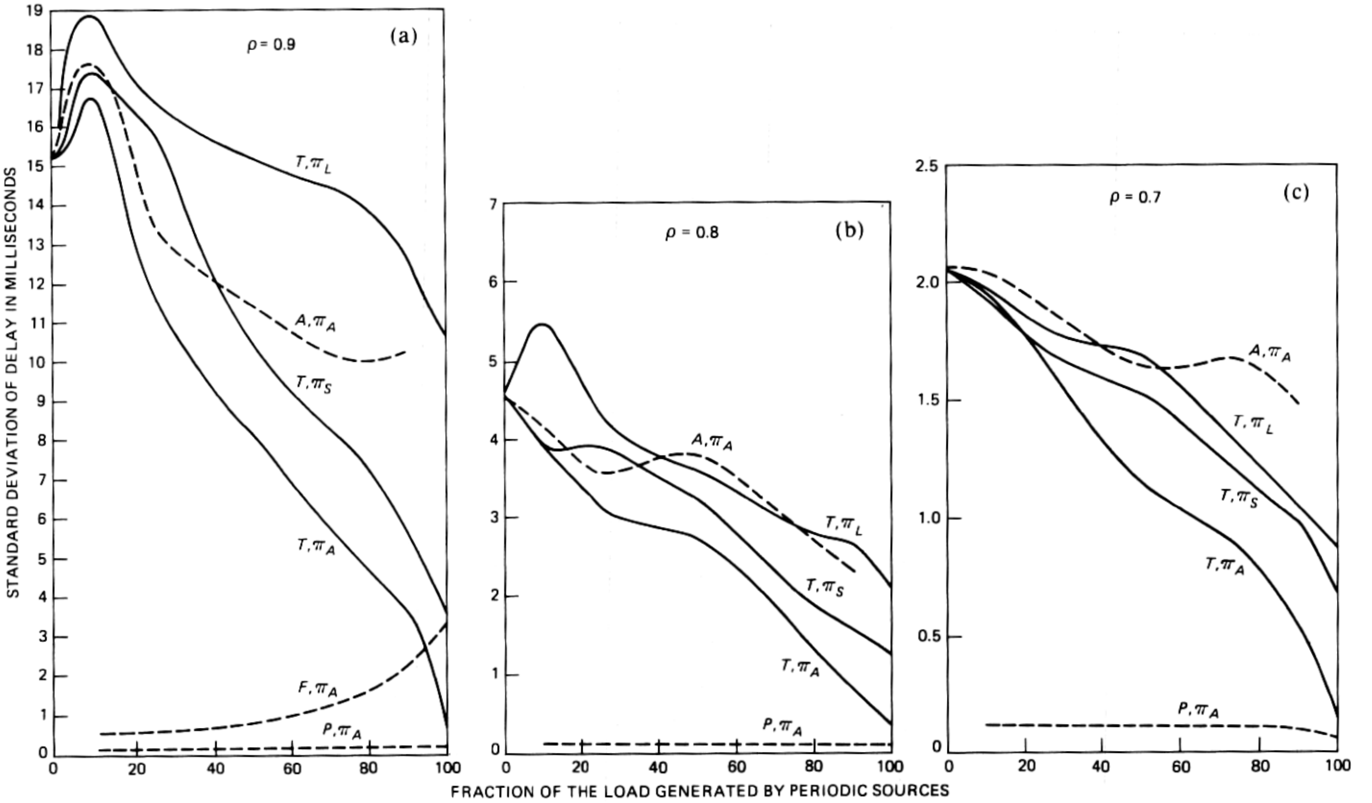


Fig. 4—Standard deviation of delay versus the fraction of the load generated by periodic sources for several types of sources with protocols at the following utilizations: (a) $\rho = 0.9$; (b) $\rho = 0.8$; (c) $\rho = 0.7$.

increases, this will comprise a larger fraction of the packet in Π_A, and the operation of Π_s will improve with respect to Π_A. A similar improvement will occur if T_p is decreased.

Protocol Π_A establishes an upper bound on the delay experienced by periodic traffic by giving this traffic a higher priority than aperiodic traffic. The dashed curves (A, Π_A) and (P, Π_A) represent the components of (T, Π_A) generated by aperiodic and periodic traffic, respectively. As expected, curves (A, Π_A) are above (T, Π_A), while (P, Π_A) are below. However, there are some interesting effects noted on these curves. As the fraction of the load generated by periodic sources increases, the number of sources that have a higher priority than the aperiodic sources increases. The delay experienced by the aperiodic sources is expected to increase. However, at utilizations of 0.8 or 0.9, there is not a significant increase in the average delay, and the standard deviation of the delay decreases.

From curves (T, Π_L), it can be concluded that it is definitely beneficial for aperiodic sources to share a channel with periodic sources. From curves (A, Π_A), it may even be beneficial for aperiodic sources to share a channel with higher priority periodic sources.

It has been stated that speech-activated systems should have a channel-acquisition delay less than 50 ms at least 98-percent of the time. From curves (A, Π_A), it should be possible to meet this constraint up to utilizations of 0.9, even if the first packet of speech is given the same transmission priority as aperiodic traffic, assuming that the 98-percent point corresponds to two standard deviations beyond the mean. In Π_A, the first packet from a periodic source is given a higher priority than packets from aperiodic sources. Using this strategy, it should be possible to lose no samples, not even those used to detect the presence of energy. For instance, if the transmitter stores the last 30 ms of speech samples, and it takes 10 ms to detect the presence of energy, no active samples are lost when the channel is acquired within 20 ms. At utilizations of 0.9, this appears to be well within the capability of the system.

X. ALTERNATIVE PROTOCOLS

Several other protocols for periodic traffic were considered. These protocols did not operate as well as the protocol described here and were not completely characterized. However, it is instructive to describe two of these protocols and their perceived shortcoming.

One possible approach to reduce the average and variance of the delay for periodic traffic is to give the periodic traffic a higher retransmission priority than aperiodic traffic. This approach is used in Π_A to reduce the delay for the first packet in a periodic sequence. Note in Figs. 3 and 4 that this strategy works best when the high-priority traffic is a very small fraction of the total traffic. Priority systems do not reduce contention between high-priority users. In a local network designed to handle data and voice requirements, the voice transmission requirements are expected to far exceed the data requirements. A priority system will not provide significant improvements in this type of an environment.

In addition, the priority system may reduce delay problems, but it does not eliminate them. The delay is still an unbounded statistical variable. The probability of exceeding a specified delay is nonzero, and packets will be lost. This probability may be acceptable for moderate utilizations, but as the utilization approaches one, it will increase and become significant. Because of the high transmission requirements of voice, it is expected that a voice and data system will have to operate at much higher utilizations than data only systems.

Another protocol, which was tried, gives periodic traffic absolute preemptive right to the network. The periodic packets have a preempt header. They begin transmitting without listening to the network and interrupt any terminals that are transmitting. The preempt header is long enough for the nonpreemptive terminals to detect a collision and stop transmitting before useful data are transmitted by the preemptive terminals. This protocol results in zero transmission delay for periodic packets once they have acquired a slot. They acquire the slot for the first packet by behaving like an aperiodic terminal, as in the recommended protocol.

This protocol requires exact timing synchronization between periodic terminals and does not operate well at high-channel utilizations. Since the periodic terminals do not listen to the channel before transmitting and transmit periodically at what they perceive to be T_p seconds, if there is a discrepancy between various terminals' estimates of T_p, they will eventually interfere with one another. In addition, when the utilization for periodic sources is slightly greater than 50 percent, it is possible for these sources to be spaced in such a way that no new sources can obtain a time slot.

When this protocol was simulated, it was found that when the channel utilization from periodic sources exceeded 0.7, the channel efficiency decreased rapidly and the access time for the first packets from periodic sources and packets from aperiodic sources became intolerably long. The channel efficiency is the fraction of time the channel is being used to transmit data that are successfully received, over the total time that data are being transmitted. The unsuccessfully transmitted data occurs when packets are interrupted by collisions. The propagation delay in local networks is small compared with the packet transmission time, and terminals listen before transmitting. Therefore, for the previously defined protocols, collisions only occur near the beginning of transmission. In this protocol, a packet can be interrupted at any time. As the fraction of time that preempt terminals use the system increases, the probability of interrupting a terminal that has completed a significant portion of its transmission increases and the channel efficiency decreases. As the channel efficiency decreases, the fraction of the channel capacity remaining for successful transmissions decreases and the delays increase. When the channel utilizations generated by periodic sources approached 0.7, the delay experienced by first periodic transmissions and aperiodic transmissions ranged from several tenths of a second to a second. This is an intolerable delay.

XI. CONCLUSIONS

When periodic and aperiodic data are transmitted on the same network, using conventional CSMA/CD protocols, the average and variance of the delay can be reduced by the periodicities in the data. By requiring periodic sources to begin generating periodic packets after they have successfully acquired the channel, the periodic requirements are separated in time, and collisions are reduced. This results in a decrease in the network delay. Channel contention is further reduced by requiring periodic sources to transmit all of the data they have acquired whenever they transmit and to schedule their next transmission a fixed time after their last successful channel acquisition.

If, in addition, periodic sources are only delayed by aperiodic sources with transmissions in progress, the size of packets from aperiodic sources is less than that from periodic sources, and the size of packets from periodic sources is not increased by the delay, then packets from periodic sources do not collide. The maximum channel acquisition delay experienced by periodic sources in this mode of operation is a packet transmission time.

By delaying the use of the first packet from a periodic source by a packet transmission time, the next packet is always received before all of the samples in previous packets are used. There are no discontinuities caused by the late arrival of samples in the reconstructed waveform, and packets of samples are never discarded because of excessive transmission delays.

In this mode of operation, periodic sources appear to have a dedicated TDM channel. There is a difference between the bandwidth allocation in this type of system and that of a true TDM system. In a TDM system, the time slots for a particular channel always carry the same number of bits and are fixed with respect to the beginning of a periodically recurring frame. In this system, the time slot for a particular channel may be moved back with respect to a hypothetical start of frame when the channel is busy. When this occurs, the number of bits of data being carried by the slot increases. In the extreme, when the channel is overutilized, the slot is shifted back by the same amount every time it occurs. The slot period and the amount of data in the slot increases.

The mobility of time slots, in conjunction with the access protocol, allows all of the excess capacity that is not being used to transmit periodic packets to be used by aperiodic sources, even if the excess capacity is dispersed in small intervals throughout the TDM frame. The system can shift between any mix of periodic and aperiodic traffic without a central controller reallocating the bandwidth.

In conventional TDM systems, all of the transmitters must be able to accurately determine the location of their assigned slots to avoid transmitting simultaneously with other sources. In this system, a periodic source is willing to wait for a communication in progress to be completed. Therefore, timing inaccuracies between periodic sources can be tolerated, as long as the inaccuracy is not long enough to cause two periodic sources to simultaneously initiate communications.

Finally, the simulation results indicate that periodic sources can acquire a channel quickly enough to operate in a time-assignment, speech-interpolation mode. Storage and delay in the system may be used to reduce, and possibly eliminate, clipping which normally occurs at the beginning of active intervals in this type of system.

REFERENCES

1. M. Schwartz, *Computer-Communication Network Design and Analysis*, New York: Prentice-Hall, 1977.
2. J. F. Hayes, "Local Distribution in Computer Communications," IEEE Communications Magazine, *19*, No. 2 (March 1981), pp. 5–14.
3. R. M. Metcalfe and D. R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Commun. of the ACM, *19*, No. 7 (July 1976), pp. 395–404.
4. R. C. Crane and E. A. Taft, "Practical Considerations in Ethernet Local Network Design," Hawaii Int. Conf. Syst. Sciences, (January 1980), pp. 166–75.
5. F. A. Tobagi, "Multiaccess Protocols in Packet Communication Systems," IEEE Trans. Commun., *COM-28*, No. 4 (April 1980), pp. 468–88.
6. T. Bailly, A. J. McLaughlin, and C. J. Weinstein, "Voice Communication in Integrated Digital Voice and Data Networks," IEEE Trans. Commun., *COM-28*, No. 9 (September 1980), pp. 1478–89.
7. T. Bailly, B. Gold, and S. Seneff, "A Technique for Adaptive Voice Flow Control in Integrated Packet Networks," IEEE Trans. Commun., *COM-28*, No. 3 (March 1980), pp. 325–33.
8. D. Cohen, "A Protocol for Packet-Switching Voice Communication," North-Holland Publishing Co., Computer Networks, *2*, No. 4–5, (September–October 1978), pp. 320–31.
9. G. J. Coviello, "Comparative Discussion of Circuit vs Packet Switched Voice," IEEE Trans. on Commun., *COM-27*, No. 8 (August 1979), pp. 1153–9.
10. O. A. Mowafi and W. J. Kelly, "Integrated Voice/Data Packet Switching Techniques for Future Military Networks," IEEE Trans. Commun., *COM-28*, No. 9 (September 1980), pp. 1655–62.
11. M. J. Fisher and T. C. Harris, "A Model for Evaluating the Performance of an Integrated Circuit- and Packet-Switched Multiplex Structure," IEEE Trans. Commun., *COM-24*, No. 2 (February 1976), pp. 195–202.
12. J. F. Shoch and J. A. Hupp, "Measured Performance of an Ethernet Local Network," Commun. of the ACM, *23*, No. 12 (December 1980), pp. 711–21.

13. J. F. Shoch, "Carrying Voice Traffic Through an Ethernet Local Network—A General Overview," IFIP WG 6.4 International Workshop on Local-Area Computer Networks, Zurich (August 1980), pp. 429–46.
14. D. H. Johnson and G. C. O'Leary, "A Local Access Network for Packetized Digital Voice Communication," IEEE Trans. Commun., *COM-29,* No. 5 (May 1981), pp. 679–88.
15. S. I. Campanella, "Digital Speech Interpolation," Comsat Tech. Rev., *6,* No. 1 (Spring 1976), pp. 127–58.
16. K. Bullington and I. M. Fraser, "Engineering Aspects of TASI," B.S.T.J., *38,* No. 2 (March 1959), pp. 353–64.
17. P. T. Brady, "A Technique for Investigating On-Off Patterns of Speech," B.S.T.J., *44,* No. 1 (January 1965), pp. 1–22.