VOICE/DATA INTEGRATION ON ETHERNET:
BACKOFF AND PRIORITY CONSIDERATIONS

by

I. Chlamtac and M. Eisengr

Technical Report #273

May 1983.
ABSTRACT

This report describes the study of the CSMA-CD protocol as given in the Ethernet specifications, for real time (voice in particular) and mixed voice/data environments. The objectives of this study are to predict the Ethernet's behavior under the current protocols, to point out ways of potential improvement within the realm of the current specifications, and to better understand the network behavior in these environments. Results are given in terms of relative performance of the various system measures given as a function of system and traffic models. It is shown that the support provided to real time voice on Ethernet strongly depends on the assumptions made with regard to the coexisting data users traffic and that with certain data traffic patterns the observed voice delay is much worse than previously reported. Finally, we show that without violating the Ethernet network specification voice service can be improved by simple adjustments of the Ethernet retransmission "backoff" algorithm.
1. INTRODUCTION

Much interest has been recently focused on real-time voice communication in local networks [4,5,6,7,9,11,13]. On the Ethernet local network in particular, the question of service quality for voice or any other real time application has been raised due to the use of the random access CSMA-CD protocol for data link control.

This protocol provides adequate service for data applications characterized by bursty arrival patterns, where protocol robustness and simplicity are traded for service level commitment. However, CSMA-CD does not guarantee bounded packet delays or packet delivery. Packets which are not successfully transmitted simply reschedule transmission following a random delay.

In case of real time voice the quality of voice reproduction and conversation interactibility depends heavily on the service provided to voice packets. The voice delay in digital voice transmission consists of the coding and decoding between analogue voice signals and Ethernet transmitted packets and of the packet transmission delay in the network. With CSMA-CD the transmission delay is not guaranteed and packets can be lost altogether due to excessive number of collisions.

As a consequence, it is interesting to know to what extent can mixed voice and data environment be supported under the CSMA-CD protocol. Several studies have addressed this question assuming Poisson arrival patterns for both voice and traffic sources [4,7].

In this paper we look at several models of combined voice and traffic environment and show that certain combinations can lead to
degraded voice service beyond that reported in the past. We then show that within the Ethernet network specification a trade-off exists between voice delay and the allowed packet loss rate. It is shown how by simple adjustment of the limit on the total number of retransmissions a better voice support can be achieved.

2. TRANSMITTING REAL TIME VOICE ON ETHERNET

Ethernet is an operational local network designed to provide cost-effective interconnection for large number of heterogeneous data users [2,8]. The CSMA-CD controlled data link is contention based and assumes that users are willing and able to make several retransmission attempts i.e., they back-off for random time intervals, prior to a successful acquisition of the channel. The packet delays are thus not bounded and depend on the prevailing traffic conditions. No guarantees on packet delivery are made. This probabilistic nature of the data link protocol is unsuitable for real-time applications. With real time voice guaranteed, and preferably constant delays and packet loss rates are among primary service objectives.

Under current industry standards the voice is digitized by sampling and coding the analog signal amplitudes once every 125 μs in 8 bit binary codes. The encoded samples are collected into packets which are transmitted over the network. The resulting data rate of the digitized voice signal is 64 K bit/sec [10]. The total delay a voice sample will endure is primarily determined by the amount of time it is queued in the packetization process and the
time it spends in transit. To reduce packetization delay (short) packet samples should be sent at short intervals. With GSMA-CD, however, the level of service, i.e., delay and loss rate are for a given traffic load, inversely proportional to the size of the submitted packets. Figure 1 shows the average packet delay for two different packet size means. The model assumes a Poisson packet arrival pattern and uniform packet size distribution. Packets are generated from a collection of 1000 (the maximum number of supportable users) transmitting over a 10 Mbit/sec channel.

Performance To what extent can integrated voice and data be supported on Ethernet can be translated to the question of how many voice users can be added to the network with existing data traffic while preserving satisfactory voice connection quality.

The specific questions we address are therefore:

- Level of support given to voice type traffic under various arrival pattern models (in terms of average delay, queue length, capacity, number of supportable users, etc.).
- The relationships between these measures and system parameters (e.g. maximum loss rate, back-off policies).
- Alternative approaches for better voice handling within the Ethernet specifications.

Most results were obtained by using a new simulation tool which provides the flexibility to configure various Ethernet system and arrival pattern models efficiently [1].

The input models for the simulation included:

- Conventional single Poisson data stream split into N (number of users) sub-streams;
- Constant, uniform, exponential and bulk exponential interarrival distributions for data packets;
- Constant rate voice packet arrivals;
- Poisson stream of voice connections (conversations) each consisting of a large number of packets arriving at uniformly spaced intervals;
- Mixed voice/data arrivals with varying ratios of the two, and with different distributions for each traffic type;
- Variable and constant messages (collection of packets) and packets size distributions.

The system was characterized by the following parameters:

- Number of nodes \( N \);
- Distribution of nodes into voice and data type nodes;
- Number of retransmission attempts and the standard CSMA-CD protocol executed by all nodes;

All collected measures represent in one way or another, the packet flow in the network. In a system with a non-negative probability of packet abortion, these measures must be taken over all generated packets or over all successfully transmitted packets as is meaningful per measure. The following measures were taken over all the generated packets:

- The average length of a node queue (i.e., number of frames awaiting transmission);
- The maximum length of a node queue;
- The total packet arrival rate;
- The channel measured throughput;
- Percentage of lost packets;
- Percentage of collided packets.
The following measures were defined for successfully transmitted packets only:

- The average packet delay per traffic class (data, voice)
- The average queueing delay (waiting time prior to the first transmission attempt)
- Percentiles of packets arriving at destination within a given time.

The traffic intensity considered is that created by new arriving packets only.

The following sections describe the various models, and provide a discussion on the results obtained for each.

3. VOICE/DATA MODELS WITH STANDARD CSMA-CD PROTOCOL

The Ethernet behavior for data only environment is significant in establishing a benchmark for later comparisons with integrated voice/data models as well as for purposes of simulation validation.

3.1 The Basic Data Models - Varying Packet length and Network Size

This is a simple homogeneous traffic data model which serves to validate the simulation programs for known traffic environments [12]. The modeled network consists of 50 nodes generating 150 and 600 byte packets of data with constant, uniform and exponential distributions around these two means. FIFO order of service is assumed at each node.

The packet header (20 bytes) are added to each generated packet prior to its first transmission attempt. For all measures—average packet delay, queueing delay and queue length, given as a function of traffic intensity, longer packets improve the system behavior as suggested in the preceding section. Results in Figure 1 correspond
closely to those provided by other models [3,4].

Figure 2 gives queueing delays - the measure least affected by the size of the packets. The difference between 150 and 600 byte packet models were not significant for queueing delay for most of the traffic intensity range as long as the system was stable.

As Figures 1-2 indicate, the system can stably support traffic intensity of up to .6 to .7 for medium to large size packets (600 bytes and up). At these rates, the delays are on the order of up to several milliseconds. The number of packets aborted after 15 retransmissions is close to zero (0.01% to 0.1%). At higher packet arrival rates (λ > .65), there is a sharp rise in the delay and the number of collided and aborted packets. Beyond (λ > .75) the average delay is by itself not a sufficient measure of system behavior since it accounts for successfully transmitted packets only. Since the delay of every successful packet is always final, the average delay is always bounded in this model.

Most existing models (analytic and simulation) assume an infinite source of users generating a poisson stream of packet arrivals [13]. In this model we want to test the finite node case. We use N = 50, 250, and 500 (number of nodes) to generate Figure 1. In all three N cases we use the previous arrival model (No.1) with the packet mean length of 150 bytes. The interarrival times are drawn from

\[ P(\lambda = \sum_{i=1}^{N} \lambda_i), \quad \lambda_i - \text{arrival rate at node } i \]

The source (node) distribution is uniform U[1,N]. The three N cases are compared for various values of arrival rates.
Figure 3 gives the average delay behavior which for low traffic intensities, $\lambda \leq .4$ is virtually identical for all $N$ values. For higher arrival rates the performance deteriorates as the number of nodes is increased. The reason is that the number of potential collision sources increases for a given rate.

In such cases for small $N$ values, packets arriving at intervals smaller than the propagation delay have a higher probability to be sampled to arrive at the same source node and thus will not collide. For small $N$'s the channel access is thus more orderly, although it requires a larger queue size at each node. These two properties become even more pronounced for small network sizes ($N < 150$).

4. VOICE ONLY MODELS

We start the investigation of real time voice on Ethernet by observing the "ideal" voice only models.

4.1 Constant packet arrivals

When voice alone is considered, theoretical upper bound on the number of voice users that can be supported simultaneously on Ethernet is obtained by assuming all voice packets arriving at constant intervals of time. Put differently we assume no voice packet collisions occur. This case is had e.g., by controlling the initial access to the channel at the beginning of each voice connection. Since subsequent packets are equally spaced, if the
beginning of the connections is properly spaced, no further collisions will occur. This upper bound is primarily a function of the vocoding rate. If we assume 64K bit/sec vocoding which generates 8 byte/sec packets every 20 msec, 160 byte packets can be sent every 20 msec. The transmission time of this packet plus headers is 144 msec. Consequently, the maximum number of users will be approx. 140. Higher numbers of users can be supported by going to a higher vocoding rate.

Figure 4 gives the maximum number of users as a function of vocoding rate at 20 and 40 msec. packetization intervals. It is especially for these faster vocoders that the current Ethernet cannot fully increase the number of users due to its inefficiency at small packet sizes.

As an example, let us consider 16K b/s coding and its transmission on Ethernet. Transmitting a packet of 80 bytes every 40 msec. after adding all headers, we get 80 msec. transmission time.

In other words, 500 is the maximum number of users, but an overhead of 40 msec. in collecting this much data via voice coding prior to transmission is incurred.

If we want to transmit every 20 msec., at this coding rate packets of only 40 bytes are generated. Taking the minimum size 46 byte packets and adding headers we find that the theoretical maximum number of connections becomes only 380! (In reality, this number will be very much lower since maximum utilization for small size packets is low.)

4.2 Exponential Connections Interarrivals

Assuming that connection (conversation) initiations come from a large collection of independent users, we can model the beginnings
of connections as coming from a poisson stream. This models a user wishing to communicate starting a connection at random. The duration of each connection is very long relative to the duration of a typical simulation run. Thus typically, new voice users join the network (until the required total arrival rate is obtained) without leaving it during the simulation run. It may be argued that after cancelling the effects of the transient phase, this model is similar to one where voice nodes dynamically join and leave the network as these events i.e., beginning and ending of a conversation, are very rare compared to its duration.

Figures 5-8 describe this voice model behavior in terms of average packet delay, queueing delay, percentage of lost packets and collisions and delay percentiles. Figure 5 gives the average delay/traffic intensity behavior. The delay distribution corresponds to the previously published results only for very low loads [4]. For higher loads the delay is significantly higher. Figure 6 gives the queueing delay with behavior similar to that of the average packet delay. The correspondence is clear since, as Figure 8 shows, for backoff = 15 the percentage of lost (aborted) packets is closed to zero even at $\lambda = .7$. Figure 7 shows percentage collisions. This value increases sharply beyond $\lambda = .3$, which explains the increase in packet delay and motivates the following model (Section 6), where the effect of decreased number of retransmissions on other performance measures is considered. Table 1 gives three different percentile values for various arrival rates for voice packet delay. Figure 8 shows the number of lost packets. It is close to zero for this standard model where the maximum number of retransmissions is 15.
5. INTEGRATED VOICE/DATA MODEL

This is the basic model of the integrating voice and data environment. The technical aspects of integration are discussed in [5,6]. To be consistent with previous models, we use the same data traffic model and the voice coding rate of 64K b/s.

5.1 Simple Poisson Arrival Model

In this arrival model a new user arrives from a poisson stream and is immediately sampled to establish himself as data or voice node. The average length of data packet is 600 bytes. Voice packets are transmitted in constant 20 msec. intervals and are of constant 160 byte length. The total arrival rate is the sum of data and voice packets. The channel throughput is obtained from simulation. The output statistics is collected separately for voice and data packets. The model includes a total of 100 nodes, of which half are voice and half are data users.

Figure 9 gives the delay/traffic performance. For low to medium loads the overall behavior is similar to that observed for voice and for data alone in other models. For high loads both voice and data degrade relatively more. It again seems that the maximum number of supportable users will be low. Table 2 gives the various percentiles as in Table 1. They are slightly worse in the "mixed" model than in the "whole only" model. While these results may be realistic for some environments, they certainly do not constitute a guaranteed level of service. As the next model shows, other data traffic types may cause a (very severe) further degradation in the average delay times for voice packets (and of course for data packets as well - a less bothering point since for data users the delays are often not of foremost importance).
5.2 "Standard" Voice Model

This model differs from the preceding one in that data arrives in long messages of 15,000 bytes each. These are transmitted in bulks of 10 packets, each of 1500 bytes - the maximum size frame for Ethernet. All measures degrade in this model and particularly the average packet delay and queueing delay grow in order of magnitude. Due to the 1-persistency of the CSMA-CD protocol a high probability of collision exists at the end of each transmission as most of the time there is a packet waiting from the same source. For example, at $\lambda = .3$, the percentage of collisions increases from 25% in the previous model to 100%. As a consequence, subsequent packets are forced to stay longer in queue before they get a chance to go on the wire. The delay experienced by voice delays is even worse since in addition to the delay caused by the overall higher percentage of collisions, the bulks of data packets awaiting transmission block voice packets for relatively long times.

As a consequence, a simple FIFO discipline at the node is not the best discipline and priority for voice should be considered. Such priority would not only enable the voice packets to get ahead, but it would do so without noticeably penalty to data since voice packets are relatively short and arrive only at constant intervals.

Figure 10 gives the delay/traffic intensity performance. Up to $\lambda = .45$ voice delays are better than data but much worse than in the conventional poisson data arrival model. Beyond this arrival rate voice packets are almost completely shut off from gaining access to channel in short (several milliseconds) time.
Figure 11 shows average and maximum queue sizes, which show a sharp increase beyond $\lambda > 0.5$. Table 3 gives packet delay percentiles which are especially worse for small delays.

6. IMPROVING VOICE/DATA INTEGRATION BY ADJUSTING THE RANGE OF ETHERNET'S BACKOFF ALGORITHM

Previous sections have investigated voice and data models when the standard data link protocol was enforced. The importance of these models lies in the attractiveness of the Ethernet local network. It is thus interesting to see if performance can be improved without basically changing this protocol. One simple step in this direction is manipulating the range of the standard backoff algorithm without changing the algorithm itself. In current standards every user will attempt to retransmit a collided packet up to fifteen times each time increasing the average backoff delay.

6.1 Voice Only Model with Reduced Number of Retransmissions

This model's objective is to study the effect of reducing the number of transmission + retransmissions from 16 to 4 and 2 on the average delay/traffic intensity loss rate measures. That a tradeoff between these measures exists can be intuitively argued as follows:

- Given a maximum allowable delay on voice packets in their decoding process, there may exist a value of retransmission beyond which a packet will arrive too late to be included in the reconstructed or retransmitted voice pattern and will thus be aborted by the voice protocol anyway.

- Given that voice decoders can tolerate a certain packet loss without significant degradation in voice quality, then by
eliminating retransmissions after a small number of attempts both average delay and throughput will be improved. The delay is improved due to limiting the maximum time a packet will spend in the node. The throughput is improved because we not only eliminate retransmissions which do not add to the effective throughput, but by doing so we also increase the probability of successful transmission for newly arrived packets.

We thus have a tradeoff situation between maximum allowable loss rate and delay/throughput behavior. Figures 5-8 and Table 1 verify these observations. Specifically, Figure 5 shows that the average delay of successfully transmitted packets decreases as the MNR (max. no. of transmission plus retransmissions) is decreased. Figure 6 shows that queueing delays decrease even more since with low MNR values each packet spends only a short time in queue. Figure 8 shows the percentage of lost packets. This figure is strongly related to Figure 7 which gives the percentage of collided packets. As seen in the case of MNR = 15, increasing traffic intensity increases sharply the number of collisions. At $\lambda \sim 0.5$ the percentage of collisions is 300%, so that a significant portion of packets becomes aborted for both MNR = 2 and 4. Between $0.25 \leq \lambda \leq 0.5$, the percentage of collisions indicates that MNR = 4 is sufficiently high to "save" most packets but MNR = 2 is not. Below $\lambda = 0.25$ any MNR is sufficient so that the maximum number of retransmissions (MNR) has no effect on packet loss behavior. This behavior is demonstrated in Figure 7 and it also explains the curves of Figure 8.

Altogether one should therefore consider reducing considerably the MNR to gain throughput and delay as shown in Figure 5 for a
given traffic intensity and allowable voice packet loss rate. For further improvement use of the backoff algorithm should be altered for voice type environments. Due to the constant arrival pattern within each voice conversation two different conversations will generate collisions on all packets given the first two packets have collided. The two possibilities to change this situation are:

- Moving the backoff in all subsequent packets in front of the packet transmissions with an effect of separating two "synchronized" voice connections, if collision on a given voice packet has been observed,

- increasing the first value of the retransmission interval to "randomize" more the constant type arrival pattern, a randomization which random access protocols of the CSMA type "prefer".

Finally, Table 1 gives three delay percentiles for voice models using MNR 2, 4 and 16. For MNR = 2 and 4, this table is only partially meaningful since for medium and high data rates the packet loss may be high for voice applications. If this loss can be sufficiently reduced, a significant delay/throughput improvement can be achieved. For MNR = 16 the first column (delay $\leq 0.26 \mu$sec.) stands for packets succeeding basically in their first transmission from an empty queue; second column (delay $\leq 2$ msec.) is the delay required by some public networks; the third column (delay $\leq 20$ msec.) is representative of the high end of the stable system throughput values.

6.2 Integrated Voice/Data with Limited Number of Voice Retransmissions

The model tests whether the previous tradeoff between delay/traffic intensity rate exists when data is present as in the previous voice only model. In the voice only case lowering the maximum number
of transmission plus retransmissions (MNR) served both the purpose of eliminating old voice packets from the system and increasing the probability of success for the new ones. The results of the integrated model support this view since not only the voice delay becomes better, but also the packet delay for data is improved even though the data packets are still allowed 15 retransmissions.

Figure 12 gives the average voice and data packet delays. The voice delay is significantly better for MNR 2 than for 16, although at the expense of increased number of aborted packets. The data delay is also better without aborting any data packets.

Figures 13 and 14 show percentages of collided and aborted packets at MNR 16, and 2. These results are similar to the voice model with limited retransmissions. Consequently, also the conclusion is similar: it is important to look for a more suitable backoff scheme to lower the number of collisions at low MNR values since the benefits in terms of delay and throughput are potentially significant. With an improved backoff scheme we can then find the optimal values of loss rate which are tolerable for voice reconstruction and thus reduce delays and increase throughput.

DISCUSSION

From the expected performance of an Ethernet local network several conclusions can be drawn:

For conventional data (e.g. timesharing) traffic the model confirms Ethernet behavior obtained from other studies. For other data arrival patterns (e.g. bulk poisson representing sudden
surge in node traffic, or short file transfers) the performance is significantly worse than reported previously. Such is also the case for voice/data integrated model for bulk data arrivals. The average delay becomes higher in an order of magnitude for most channel loads, as compared to the standard data models.

In other words, with certain data arrival patterns, both voice and data packet response times may be negatively affected. We also show that voice delay performance can not be improved by simply increasing the vocoding rate or by reducing the maximum number of retransmissions. In the case of higher vocoding rates, it is the limit on the minimum size packet which prevents us from fully utilizing this vocating trend. In the case of reducing retransmission attempts, the existing backoff algorithm can lead to untolerable loss rates at higher loads.

The results suggest that an improvement in Ethernet behavior (with slight protocol changes) can be had along the lines of properly adjusting the backoff policy to reduce packet losses and by introducing class priorities (possibly at the client layer) to reduce the negative effects of data traffic on voice and other delay constrained applications.
Figure 1: Packet delay/traffic intensity
150, 600 bytes packets

Figure 2: Waiting time in queue
vs. traffic intensity
Figure 3: Packet delay/traffic intensity for varying nbr. of nodes

Figure 4: Number of supportable users per vocoding speed (kb/s).
The ideal case for voice.
Backoff policy with different Maximum Number of Retransmissions for voice packets (MNR)

Figure 5: Average delay/traffic intensity (voice only model)

Figure 6: Queueing delay/traffic intensity
Figure 7: Packet collision percentage vs. traffic intensity

Figure 8: Percentage of lost packets vs. traffic intensity
Figure 9: Average delay vs. traffic intensity

Figure 10: Average delay vs. traffic intensity
Figure 11: Average queue size (at all nodes combined) vs. traffic intensity

![Graph showing average queue size vs. traffic intensity for voice/data bulk model, bulk data, and standard data.]

Traffic intensity

Figure 12: Voice/data integrated model. Average delay vs. traffic intensity

![Graph showing average delay vs. traffic intensity for standard backoff, reduced MNR(2), data, and voice.]

Traffic intensity
Figure 13: Percentage of collided packets vs. traffic intensity

Figure 14: Percentage of lost packets vs. traffic intensity
### TABLE 1: Percentile of Voice Packets under Given Delay: Voice Only Model

<table>
<thead>
<tr>
<th>λ</th>
<th>Packet Delay (msec)</th>
<th>0.26 msec</th>
<th>2 msec</th>
<th>20 msec</th>
</tr>
</thead>
<tbody>
<tr>
<td>.17</td>
<td>71%</td>
<td>93%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.35</td>
<td>71%</td>
<td>93%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.53</td>
<td>51%</td>
<td>87%</td>
<td>99%</td>
<td></td>
</tr>
<tr>
<td>.72</td>
<td>47%</td>
<td>82%</td>
<td>97%</td>
<td></td>
</tr>
<tr>
<td>.31</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.40</td>
<td>99%</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.49</td>
<td>97%</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.28</td>
<td>86%</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.40</td>
<td>84%</td>
<td>100%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.52</td>
<td>76%</td>
<td>99%</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

Backoff = 15

### TABLE 2: Percentiles of Voice Packets under Given Delay: Standard Voice/Data Model

<table>
<thead>
<tr>
<th>λ</th>
<th>Packet Delay (msec)</th>
<th>0.26 msec</th>
<th>2 msec</th>
<th>20 msec</th>
</tr>
</thead>
<tbody>
<tr>
<td>.24</td>
<td>.64%</td>
<td>99%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.40</td>
<td>54%</td>
<td>90%</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>.61</td>
<td>34%</td>
<td>87%</td>
<td>99%</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 3: Percentile of Voice Packets under Given Delay: Bulk Data/Voice Model

<table>
<thead>
<tr>
<th>λ</th>
<th>0.26 msec</th>
<th>2 msec</th>
<th>20 msec</th>
</tr>
</thead>
<tbody>
<tr>
<td>.25</td>
<td>49%</td>
<td>92%</td>
<td>99%</td>
</tr>
<tr>
<td>.30</td>
<td>47%</td>
<td>90%</td>
<td>98%</td>
</tr>
<tr>
<td>.50</td>
<td>32%</td>
<td>84%</td>
<td>96%</td>
</tr>
</tbody>
</table>

TABLE 4: Percentile of Voice Packets under Given Delay: Standard Data/Voice with Reduced Backoff Model (backoff = 2)

<table>
<thead>
<tr>
<th>λ</th>
<th>0.26 msec</th>
<th>2 msec</th>
<th>20 msec</th>
</tr>
</thead>
<tbody>
<tr>
<td>.23</td>
<td>80%</td>
<td>99.8%</td>
<td>100%</td>
</tr>
<tr>
<td>.30</td>
<td>75%</td>
<td>99.8%</td>
<td>100%</td>
</tr>
<tr>
<td>.75</td>
<td>54%</td>
<td>97%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Comment: At high arrival rates percentages are not significant due to high loss rate.
REFERENCES


