Applying MSTDM to Packet Voice and Data in Cellular Radio Systems

N. F. Maxemchuk
AT&T Labs - Research
Florham Park, N.J.

Abstract

MSTDM is a variant of the Ethernet protocol that provides quality of service guarantees for voice. This protocol has been applied to LAN's and CATV networks. Cellular networks are different:

1) Each cell has a base station: When mobile units communicate through the base station, frequency reuse increases and collisions are detected.

2) The objective is to connect mobile units to the communication infrastructure: Communications between units is only used to share the channel.

3) The base station broadcasts packets from the communications infrastructure to the mobile units: This channel can also provide timing and control.

Rapid contention resolution techniques make it unnecessary to transmit voice during silent intervals. Taking advantage of the cellular characteristics and TASI increases the number of voice connections by a factor of 8.

1. Introduction

The number of users in cellular networks is increasing. In order to support the larger number of users with the same bandwidth, cell sizes must decrease and the bandwidth must be reused more often. As cell sizes decrease, the propagation delay between mobile units that contend for a shared channel becomes less than the packet transmission time. Random access mechanisms that sense the channel, such as the Ethernet LAN protocols, become efficient.

In this work I describe contention mechanisms that obtain the quality of service guarantees needed for voice communications while simultaneously freeing up bandwidth and providing access for data applications. The techniques are variants of movable slot TDM (MSTDM)[1,2], which is a modified version of the carrier sense multiple access protocol with collision detection (CSMA/CD)[3] that is used on the Ethernet[4]. In the sequel[5] I show how this technique is combined with other packet switching techniques to reduce the work involved in handing off between cells and to reduce the probability of dropping a connection when entering an over populated cell.

Carrier sense multiple access protocols (CSMA)[6,7] are preferred to CSMA/CD in radio networks because collision detection is difficult to implement. Collision detection is routinely performed on properly terminated cable networks, but the reflection of a mobile unit's own signal from a nearby obstacle may be stronger than the signal from a distant transmitter in a radio network. Several of the mechanisms used in MSTDM have been used to support voice traffic in wireless LAN's[8]. Instead of performing collision detection, contending stations send a signal burst before sending a packet, then stop transmitting to determine if any other stations are transmitting. Priorities are set by varying the burst size and give preference to voice stations that have experienced the greatest delay. The contention resolution period is greater and this protocol does not provide the quality of service guarantees as MSTDM with collision detection.

Collision detection can be used in a cellular radio network by applying a two-channel technique that was developed for CATV networks[9,10]. Directional taps in a CATV system make it impossible for stations to directly receive the signal from all of the other stations. The problem is solved by receiving the "up stream" signal from all of the stations at the head-end of the network, and retransmitting it in a "down stream" channel that can be received by all of the stations. The stations transmit in the upstream channel but perform CSMA/CD on the downstream channel.

The same strategy can be used in a cellular network, with a base station instead of the head-end. The mobile units transmit in a frequency band that is received by the base station, the contention channel. The base station retransmits the signals that it receives on the contention channel in a second frequency band, the reflection channel, that is received by the mobile units. A mobile unit detects the absence of signal on the reflection channel before transmitting on the contention channel. It then listens to the reflection channel while transmitting to determine that the base station is receiving its signal without collision. In section 2 I show that a CSMA/CD system that uses two channels uses less bandwidth per unit of surface area than a CSMA system that uses one channel.

The primary objective of the cellular system is to connect mobile units with the communications network, rather than with mobile units in the same cell. Therefore, the reflection channel must only carry the state of, and not the data on, the contention channel. The amount of state information that must be carried on the reflection channel depends upon the protocol. Different information is needed to implement CSMA, CSMA/CD, MSTDM, or the variants of MSTDM that are described here.

A base station must also broadcast information from the network to the mobile units in its cell. The state information about the contention channel is not transmitted on a separate reflection channel, but is multiplexed on the broadcast channel. In sections 4 and 5 asynchronous and minislotted systems are described. Different multiplexing techniques are used in each of these systems.

Asynchronous MSTDM is derived from the earlier MSTDM systems. However, reduced state information and the variable delay to receive the state information reduces the guarantees for voice and the efficiency of data access. In the minislotted system, the voice guarantees and data efficiency are recovered. In addition, in the minislotted system the packet overhead is reduced and channel contention is resolved more quickly.

1.1 MSTDM

MSTDM is a variant of CSMA/CD that is used to transmit voice and data. With this protocol, a voice packet is delayed by less than a packet transmission time, two continuing voice sources never collide, and voice packets are never lost because of contention. These properties hold even as the network utilization approaches one.

In MSTDM the first voice packet establishes a position on the channel and the packets that follow it are given priority over contending sources. In the protocol operation:

- Data packets and the first voice packet use the conventional CSMA/CD protocol and constrain their packet size to be less than or equal to a voice packet, T_X.
- Priority voice packets use a CSMA protocol, but win collisions with non-priority packets because of a preempt interval at the beginning of a packet.
- During the preempt interval a priority source sends signal, but no information, so that other sources may detect a collision and stop transmitting before this source starts to send information.
- Successive voice packets are scheduled a fixed period T_P after the source successfully acquires the channel.
- If the channel is busy when the priority source is ready to transmit, the source transmits as soon as the channel becomes idle.
- If a priority voice source is delayed by a transmission that is in progress, any extra samples that arrive are included in an overflow area that is normally empty.

A voice source expects to acquire the channel each T_P seconds. When it is delayed by a packet in progress, the entire future sequence is delayed by this amount. Hence the name, movable slot TDM.

The voice guarantees of MSTDM are proven in reference 1. The guarantees are obtained because:

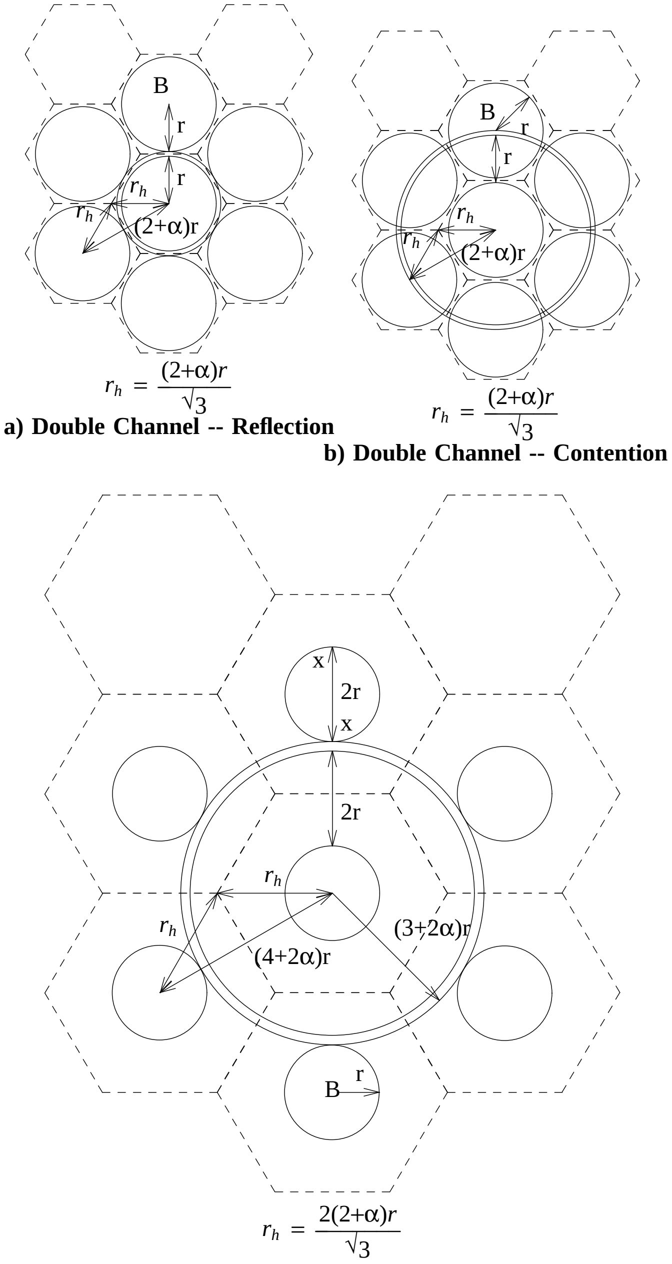
- The minimum separation between the scheduled priority arrivals is T_X since packets are scheduled to arrive T_P after the channel is acquired and each packet takes T_X to transmit.
- A priority packet can only be delayed by a non-priority packet that starts transmitting before its scheduled arrival. The delay is less than T_X, and will not force two priority packets to collide.
- A priority packet delays the next priority packet by an amount that is less than or equal to the amount that it has been delayed.

MSTDM can be implemented in a radio network by using a two channel approach that has been used in CATV networks[9,10]. However, rather than imitating the wired environment, the remainder of this work considers variants of MSTDM that can provide a similar quality of service for voice but require less bandwidth.

2. CSMA in packet radio networks

In a packet radio network mobile units listen to a channel to determine if it is busy before transmitting. A mobile unit must receive the signal that is transmitted by all of the other units in its transmission area. If the transmission area is a circle of radius r, then the signal transmitted by a mobile unit must reach receivers that are $2r$ away. A receiver in another transmission area that uses the same frequency band must not detect the signal or it will mistakenly assume that the band is in use in its own area. Therefore, two circles that use the same band must be further than $2r$ away at their closest point. Figure 1 depicts the maximum packing of circles, that can reuse the same frequency band. There is a guard band that is proportional to the transmission distance between each circle.

In a cellular system the base station can retransmit all of the signals that are received on the contention


B
r
r
r_h
r_h
(2+α)r
$r_h = \frac{(2+\alpha)r}{\sqrt{3}}$
B
r
r
r_h
r_h
(2+α)r
$r_h = \frac{(2+\alpha)r}{\sqrt{3}}$
x
2r
x
2r
r_h
r_h
(3+2α)r
(4+2α)r
B
r
$r_h = \frac{2(2+\alpha)r}{\sqrt{3}}$

a) Double Channel -- Reflection

b) Double Channel -- Contention

Single Channel

Figure 1. Frequency Reuse in Single and Double Channel CSMA Networks

channel on a second, reflection channel. In the two channel system a mobile unit must only transmit sufficient power to reach a base station in the center of the circle. A second base station, that receives on the same frequency, must be far enough away that it does not receive the signal from this mobile unit. The packing of transmission areas that meet this constraint is depicted as the contention channel in figure 1. Once again, the guard band is proportional to the transmission distance.

The base station must transmit sufficient power to be received by every mobile unit in its transmission area, but not be by mobile units in adjacent areas. In figure 1 the reflection channel depicts the packing of transmission areas that satisfy this constraint. The constraints on the contention and reflection channels result in the same packing.

The centers of the circles in figure 1 are arranged in a hexagonal array so that the distance between all neighboring circles is the same. This geometry provides the densest packing of circles. The line between the center of each circle is bisected to form the dashed hexagons. The circle is the area that uses the frequency band for communications and the hexagon is the average area in which the frequency band cannot be reused. The ratio is the fraction of an infinite surface that uses a frequency band for communications.

In the two channel system the ratio of the area of the circle to the area of the encompassing hexagons is $R_2 = \frac{2\pi}{\sqrt{3}(2+\alpha)^2}$. In the single channel system the distance between the edges of adjacent circles is $2(r+\alpha)$, and the ratio of the areas is $R_1 = 4*R_2$. Using a frequency band in 4 times as many transmission areas, makes four times as many bands available in each area.

When the surface is subdivided into equal size transmission areas it may not be possible to place circles on the exact centers that are dictated by the constraints. In effect, frequency bands are not reused as closely as they can be. In the sequel[5] frequency plans are described for the constraints resulting from single and double channel systems. The frequency plans make 7/3, instead of 4 times, as much bandwidth available in each cell in the two channel system as in the one channel system.

3. Two channel systems

In the two channel systems the reflection channel is not used to communicate between sources, but only to determine the state of the contention channel. Therefore, the reflection channel does not require the same bandwidth as the contention channel. There is a trade off between the amount of state information that is reflected and how well contention resolution can be performed. In the remaining sections we consider two systems, an asynchronous and a slotted system. In the asynchronous system a mobile unit may contend for an idle channel at any time. In a slotted system, the time following a transmission is partitioned into minislots by the base station. Both systems use only a few channel states, but their operation is very different.

Before describing the differences between asynchronous and slotted systems, I will describe the common characteristics. The objective in both systems is to communicate between a single base station and many mobile units in a cell. The base station connects the mobile unit to the communications infrastructure. In addition to the contention and reflection channels, there is a third channel, the broadcast channel, that is used by the base station to send packets from the infrastructure to the mobile units. The base station is the only source on the broadcast channel and transmits continuously.

The state information that is carried on the reflection channel is multiplexed onto the broadcast channel, instead of using a separate frequency band. The multiplexing technique is different for the asynchronous and slotted system, but both techniques make it possible to report state changes in a timely manner.

The systems support both data and voice, and use variants of the MSTDM protocol. Bandwidth is the scarce resource in cellular radio networks. Since speech conversations are only active 40% of the time[11], a TASI[12] mode of operation is used, and transmission is suppressed during silent intervals.. TASI is simpler to implement in MSTDM than in a circuit switched system because the unused bandwidth is automatically available for contention by other sources when a channel becomes idle.

A problem with contention based mechanisms is that there may be a delay before the first packet in an active voice interval can establish a priority sequence. In TASI systems the total number of connections is constrained so that 95% of the connections that become active can be assigned a circuit immediately. The remaining 5% of the active intervals are clipped until a circuit becomes available. This constraint can be translated into the contention domain by requiring that 95% of the first speech packets acquire the channel within a packet assembly time.

There are three reasons to be less concerned with this limit in a cellular network than in the original TASI networks:

1. Reduced storage costs makes it possible to delay, rather than clip active intervals. (Delaying an active interval compresses the next silent interval.)
2. Typical MSTDM networks are designed to handle large numbers of voice connections so that the variance in the number of active users is small.
3. The cellular network is designed for both data and voice so that the access delay can be reduced by giving the initial voice packets priority over data.

In the 60's storage was relatively expensive and TASI circuits were designed to clip rather than store speech. By the late 70's TASI systems were proposed that traded clipping for delay[13]. When speech is delayed rather than clipped the constraint on access delay can be relaxed from a packet assembly time to a fraction of a silent interval. Packet assembly times are 20 to 50 msecs. while the average silent interval is about 1.7 seconds. Therefore, changing the criterion can increase the deadline for transmitting the first packet in an active interval by up to an order of magnitude.

The number of voice connections that can be supported by MSTDM depends upon the bit rate of the system, the rate of the coder, the packet assembly interval and the size of the packet headers. These relationships are investigated in reference [2]. For a 10 Mbps Ethernet with 32 kbps voice coders, the number of voice users that the system can support is between 150 and 300. Therefore, MSTDM systems are designed for a large numbers of users.

When there are n users talking independently, the number of active users, n_a, is a binomial distribution with mean $.4n$ and variance $.24n$. Since n is large, the binomial distribution is approximated by a normal distribution with the same mean and variance, $N(.4n,.24n)$. The number of active voice users exceeds $F_a(95) = .4n+1.645\sqrt{.24n}$ 5% of the time and exceeds $F_a(99) = .4n+3\sqrt{.24n}$ 1% of the time. When n=100, $F_a(95) = 48$, and $F_a(99) = 55$. Therefore, the TASI quality of service can be guaranteed if new voice connections almost always succeed in establishing a connection when there are 48 active connections. A better quality is obtained if access can be guaranteed when there are 55 active users.

In the original TASI systems the penalty for guaranteeing access when there are more active users is a decrease in system utilization. For example, when $n = 100$ the average number of active users is 40, therefore when we plan for a maximum of 48 users the channel utilization is about 83%, while if we plan for 55 users the utilization drops to about 73%. The penalty is not the same in the cellular network. We assume that the objective is to carry data and voice and that the bandwidth that is not currently assigned to voice will be contended for by the data sources. Leaving at least 25% of the bandwidth for data contention is not excessive.

In the TASI system the model for the users that become active, n_f, is a Poisson distribution with an arrival rate per second of $n_s/1.7$, where $n_s = n - n_a$ is the number of silent users. The arrival rate in an interval of T *secs* is $\lambda = T*(n - n_a)/1.7$. Approximating the Poisson distribution as $N(\lambda,\lambda)$, when $n_a = 55$ and $T = .05$, the packet assembly time, $\overline{n_f} = 1.32$ and $F_f(99) = 5$. Therefore, if contention between 5 sources can be resolved during a packet assembly period then the first packet of an active interval is delayed more than the packet assembly time less than 1% of the time.

4. Asynchronous protocol

The asynchronous protocol is a modified version of the original MSTDM protocol. It uses three states of the contention channel, as opposed to the more complete information that is needed to identify collisions in a CSMA/CD protocol. The channel states indicate that the channel is idle, S_I, busy carrying a low priority packet, S_D, or busy carrying a high priority, voice packet, S_V. The channel states change asynchronously with respect to the packets that are transmitted on the broadcast channel and the changes are reported using a data link escape sequence (DLE).

Low and high priority packets are identified by the timing sequence at the beginning of the packet. The timing sequence at the beginning of a packet is long enough for the base station to detect the presence of the packet and to extract symbol timing. Low priority packets have a periodic signal at the symbol rate and high priority packets at half that rate. If multiple sources collide, nonlinear components are likely to form sum and difference frequencies. If several low priority sources collide, they are not mistaken for a high priority source since the half frequency is never be formed. Whenever a high priority packet is transmitted there is an appreciable component at half the symbol rate even though low priority sources are transmitting.

When the state of the contention channel changes, the base station must report the change promptly in order for the information to be useful for preventing and detecting collisions. The base station interrupts any transmission on the broadcast channel with a DLE and reports the state of the contention channel.

DLE's were used in the earliest computer communications protocols and can be either bit or byte oriented. A bit oriented data link escape is used in IBM's synchronous data link communication (SDLC) protocol. A sequence of seven consecutive one's is inserted in the data prior to transmitting a control sequence. A zero is inserted in the transmitted data whenever six ones occur, whether or not the next bit is already zero, to prevent the occurrence of seven one's. When six one's followed by a zero is received, the zero is deleted. When seven one's are received, all seven one's and the control character are removed from the data. Early byte oriented DLE protocols included IBM's binary synchronous communications (BSC) and ANSI's data communication control procedures 1, (DCCP-1). A one byte DLE character precedes any control character. If DLE occurs in the data stream, the control character that follows it specifies that the DLE is genuine.

In general, the bit oriented protocol reports a change in channel state more quickly. When the base station detects a change in the carrier it immediately starts transmitting one's followed by a two bit control sequence. The change in channel state is reported in 9 bit transmission times. In the byte oriented protocol the base station waits for the next byte boundary, then transmits the DLE and control byte, which can take between 16 and 24 bit transmission times. Because general purpose processors are byte oriented, byte oriented protocols are less expensive to implement. In the remainder of this work we assume a byte oriented DLE protocol.

Data packets and the first packet in a voice sequence are low priority. They transmit if the channel is idle and continue to transmit if the channel state does not change to S_V. This is different than the original MSTDM protocol since low priority packets detect collisions with high priority packets but not other low priority packets. The packets contain a redundancy check and the base station acknowledges packets that are successfully received. Packets that are not acknowledged are retransmitted as low priority packets. When a high priority voice packet is retransmitted, the retrys only continue until the packet becomes stale. The packet is stale when the decoder has passed the voice samples that it contains.

In effect, low priority packets use a CSMA/CD protocol to contend with high priority packets and a CSMA protocol to contend with other low priority packets. High priority voice packets use a CSMA protocol. CSMA/CD is more efficient than CSMA, however,

1. as the propagation delay in a cellular system decreases, the performance differences between CSMA and CSMA/CD also decrease, and
2. the efficiency of MSTDM is greater than CSMA or CSMA/CD because the fraction of the bandwidth that supports high priority sources does not go unused during contention.

In the original version of MSTDM the fixed period, T_P, between acquiring the channel and the next channel acquisition is sufficient to guarantee that two high priority packets never collide. The proof depends upon high priority packets not being delayed more than the packets that precede them. In the current implementation the base station does not immediately report a change in channel state, but waits for a byte boundary. Waiting for a byte boundary can increase the delay of successive voice sources. There is small probability that the increase in delay will cause two preemptive sources to collide. The protocol addresses a collision between two preemptive sources by not acknowledging packets that collide.

Asynchronous MSTDM does not have the strict quality of service guarantees for voice nor the efficiency of CSMA/CD of the original MSTDM. Even with the degradation, the guarantees and efficiency are better than those provided by standard channel contention techniques, such as CSMA.

4.1 Operation

When a high priority voice source encounters a busy channel it transmits as soon as the channel is idle. When the first packet from a voice source encounters a busy channel, or does not receive an acknowledgement, it waits T_V from the time that the channel becomes idle before retransmitting. T_V is a random variable with minimum and maximum values $T_{V,min}$ and $T_{V,max}$. A data source operates the same way that the first packet from a voice source operates, except that is waits T_D from a distribution with $[T_{D,min}, T_{D,max}]$.

Mobile units sense that the channel is idle when they receive the control message from the base station. The maximum difference in times when mobile units detect an idle channel is $\rho_{\max}$, the propagation delay from the base station to the furthest mobile unit in the cell.

When $T_{D,min} > T_{V,max} + \rho_{\max}$, the data source only contends for the channel when there are no new voice sources trying to acquire bandwidth. With this priority mechanism, voice sources can suppress silent intervals because they can quickly acquire the channel. Data sources only contend for the channel when the voice requirements so that the degree of oversubscription for voice sources is the same as in voice only systems[13.

4.2 Timing intervals

The timing sequence at the beginning of a voice packet is long enough to allow data terminals that are transmitting to stop interfering with the voice signal. The sequence at the beginning of a data packet is long enough that the clock at the beginning of a voice packet can be detected before the data in the data packet is transmitted and can introduce a half frequency component.

Figure 2 shows the contention interval at the beginning of a transmission. In this figure:

— ρ_V and ρ_D are the propagation times between the base station and a voice and data terminal,

— Δ_E and Δ_V are the times needed to reliably detect a carrier in the transmission band and the half frequency component that identifies a preemptive voice source, and

— W_B and X_C are the times required to wait for a byte boundary in the base station's transmit channel and the time required to transmit a control sequence that signifies the state of the transmit channel.

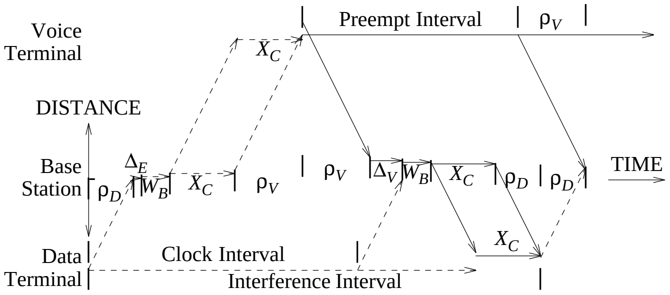


Figure 2. Timing diagram for Competing Stations in an Asynchronous System

The preempt interval at the beginning of a voice packet must be $\tau_{PRI} = \Delta_V + W_B + X_C + 2\rho_{\max}$ in order for lower priority units to stop transmitting without interfering with the voice station's payload. ($\rho_{\max}$ is the maximum propagation time between a base station and a mobile unit.) A data packet transmits the clock frequency for at least $\tau_{CI} = \Delta_E + \tau_{PRI}$ before transmitting data, which is long enough to guarantee that the base station receives the clock from a competing voice station. The data station must transmit its own clock during this interval because a data signal may have a component that is misinterpreted as the voice clock. Since the intervals are similar in length the same size clock interval is be used for all packets.

The propagation delay of light is about 3 1⁄3 bit transmission times for each kilometer it travels and each Megabit per second that is transmitted. Therefore, in a cellular network with 1/2 kilometer cells and 10 Mbps in each frequency band, $\rho_{\max} = 16\ 2/3$ bits. In a system with a high signal to noise ratio, it should be possible to detect carrier in about a symbol transmission time and extract clock 1 or 2 symbols after the signal is detected. In a system that transmits 4 bits per symbol Δ_e and Δ_V are about than 4 and 12 bits. In addition, $W_B < 8$ bits and $X_C = 16$ bits. Therefore, in this sample system, the clock interval, or preempt interval, at the beginning of a packet should be about 8 1⁄2 bytes. The interference interval at the beginning of a packet, during which a second source may start transmitting and collide, is 16 bytes.

5. Slotted system

An important difference between the cellular system and most earlier MSTDM implementations is that there is a channel from the base station to all of the mobile units that is transmitting continuously. The signal that is received from the base station can provide bit timing for the contention channel. By placing a little more structure in the signal it can also divide the contention interval into minislots. The structure can be obtained by transmitting single bits, that follow a known sequence, on the broadcast channel to mark each minislot.

The minislots are long enough for the beginning of slot signal to reach any mobile unit in the cell, for that unit to elect to transmit or not and for the signal from the mobile unit to reach and be detected by the base station, as shown in figure 3. Using the the values for propagation delay and timing extraction that were used in the asynchronous system, and a 4 bit control sequence, the minimum size minislot is about 5 bytes, which is less than the 8 1/2 byte preempt interval and 16 byte interference interval in the asynchronous system. This is a more efficient method for contending for a channel[6,14,15], since packets only collide if they start transmitting in the same minislot.

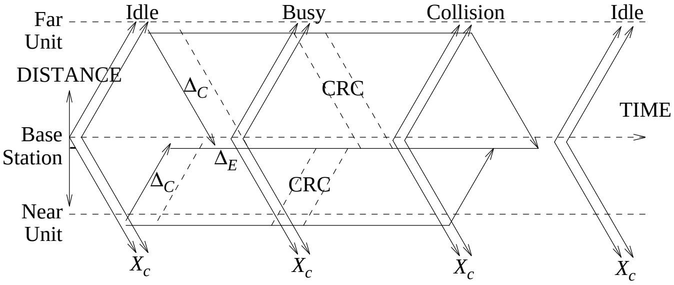


Figure 3. Timing diagram for A Collision In a Slotted System

A minislotted system with arbitrary size packets can be implemented by starting minislots at the end of a packet transmission at a base station. However, to simplify some other improvements, and the proof for voice quality of service, voice packets are n_X minislots, data packets are $\leq n_X$ minislots, and the time for the next scheduled packet from an active voice source is n_P minislots after the beginning of transmission of the previous voice packet.

The state of the contention channel is reported periodically when the base station transmits the bit sequence that marks the beginning of a minislot. As a result, it is not necessary to transmit the DLE before a control sequence. The penalty is that the control bits are transmitted once per minislot instead of twice per packet. There are operating regions where each approach sends more control bits than the other, but the difference is not significant.

5.1 Improvements in slotted systems

Once the contention channel is slotted and the state is reported periodically it is also possible to make the following modifications to improve the operation of the protocol:

1. Track the expected voice sources at the base station and use the channel state bits to prevent collisions in that minislot.
2. Detect collisions between sources and use the channel state bits to stop colliding sources.
3 Organize the contention resolution algorithm as a tree search.

5.1.1 Collision Prevention. The packet accumulation period for voice is a fixed number, n_P, minislots. The base station expects a voice source to be ready to transmit its next packet n_P minislots after its last transmission. If the contention channel is idle, the base station changes the state of the next minislot to indicate that it is reserved for voice. If the contention channel is busy, the base station waits until it becomes idle and then reserves the contention channel for voice. Once again, if a voice source is delayed it sends the extra samples that arrive in an overflow area so that it does not transmit again until n_P minislots after it acquired the channel.

The base station sends the reservation for a voice source for one minislot. If the source accepts the reservation the channel becomes busy. When a voice source becomes silent the reserved minislot returns empty and the reservation is removed from the base station's list. The next minislot is either an invitation to low priority sources or an invitation to the next reserved voice source. (The latter condition only occurs when the previous reservation is delayed by $n_x - 1$ minislots.)

There is no need for a preempt header on the voice packet, since a low priority source will not transmit in a minislot that is reserved for voice. The penalty is that a minislot is wasted when a voice source becomes silent. The average active interval is about 1.3 seconds and the packet accumulation period is less than 50 msecs. Therefore, there are more than 25 packets per active interval and the slotted system uses less than 4% as many bits to establish priority as the asynchronous system.

With this adaptation the slotted system completely restores the quality of voice service of the original MSTDM.

Scheduled sources are not be delayed by scheduled sources that were not delayed. Reservations for successive voice packets are at least n_X minislots apart, since each voice source receives a reservation n_P slots after its last reservation and each voice transmission lasts n_X minislots.

Therefore, the first scheduled source that is delayed must be delayed by a low priority source.

A low priority source delays a scheduled source by $n_X - 1$ or less slots. When a source is scheduled to transmit the minislot is reserved and a low priority source cannot transmit. Therefore, a scheduled source can only be delayed by a low priority source that has already acquired the channel, and the packet size of a low priority source is $\leq n_X$.

A scheduled source that is waiting for the channel is not delayed by a low priority source. When a scheduled source is waiting and the channel becomes idle, the first minislot is reserved for the scheduled source.

Therefore, low priority sources delay scheduled sources $< n_X$ and cannot force two scheduled source to collide since they are scheduled $\geq n_X$ minislots apart.

A scheduled source that is delayed delays the next scheduled source by the number of minislots that it was delayed or less. Scheduled sources transmit for n_X minislots and are scheduled $\geq n_X$ minislots apart.

Therefore, a scheduled source that is delayed $< n_X$ cannot delay the next scheduled source by $\geq n_X$ and force it to collide with a subsequent scheduled source.

As a result, two scheduled voice sources never collide.

Slotted reservation systems have been used for voice access in packet radio[16]. When there are a fixed number of voice channels, MSTDM and slotted reservations are the same. However, MSTDM supports variable packet sizes for data and resolves contention for new voice sources more quickly than the earlier proposals. Particularly when using the contention resolution algorithms in section 5.1.3. The faster resolution means that TASI operation, where packets are not sent during silent intervals, can be sent at higher channel utilizations.

5.1.2 Collision Detection. While the channel is idle the base station transmits invitations to transmit in each minislot. When one or more sources start transmitting the base station indicates that the next minislot is under contention, and no additional sources start transmitting in this slot. In the first minislot a source includes a redundancy check. The redundancy check is received by the base station before it transmits the state of the of the third minislot.

If the redundancy checks, the base station makes the third minislot busy and the mobile unit continues to transmit its packet. If the redundancy does not check the base station sends a collision notification in the third slot and the fourth slot will be an invitation to the next voice source or an invitation to contend for the channel.

Collisions are detected by the redundancy check and a collision lasts three minislots, rather than a packet time. This is on the same order of time that is required to act on collisions in CSMA/CD networks. Whether or not the savings are significant depends upon the size of a packet relative to a minislot.

5.1.3 Contention Resolution. Minislotted systems lend themselves to contention resolution algorithms that are based on a tree search[17,18]. These algorithms resolve conflicts more quickly by reducing the number of users that are contending for the channel. They also provide a means of assigning priority to the first packet in a voice sequence over a data packet.

When there is a collision in a slot the contenders are divided into two sets, the "0" set that will contend for the next slot and the "1" set that waits until all contention in the "0" set is resolved. If there is a collision in the "0" set's slot, the set it is divided into a "00" set that continues to contend and a "01" set that waits until the "00" set completes transmission. A set completes transmission if a source successfully transmits a packet or if the slot is empty. The "0" set completes transmission when both the "00" and "01" sets complete transmission.

In a simple implementation, when a source enters the "wait" state, the number of sets it is waiting for, S_W, is set to 1. Each time there is a collision S_W is increased by one and each time a set completes S_W is decreased by 1. When $S_W = 0$ it is the waiting source's turn to transmit. There are modifications that reduce the number of steps that are needed to resolve contention, see reference 19, page 292.

The sets can be assigned by flipping a coin, by a unique number such as the source's address, or by the time that the packet arrived. To give the first packet in a voice sequence priority over a data source, following the first collision all data sources join set "1" and all voice sources join set "0". When the remaining steps in the contention resolution algorithm are based upon the arrival time, the variance of delay is reduced. When the remaining steps in the contention resolution algorithm are based upon the source address, the delay is bounded.

Since this is a mobile network the contention resolution algorithm must accommodate sources that enter the cell and are not tracking the sets. This can be accomplished with an "unblocked stack algorithm" in which a new source joins the set that can transmit. In order to maintain voice priority, the base station transmits two idle channel indicators, the first indicates that the idle slot is available for voice only and the second indicates that it is available for any source. A voice source that becomes active joins any set that is contending and a data source waits until the idle slot is available for any source before doing the same. Regularly, when there is a long contention period for data sources, the base station sets aside a minislot for voice terminals, in case any voice sources have become involved in the channel contention.

5.2 Implementation

An important difference between slotted and asynchronous systems is that the base station operates as a controller rather than a simple regenerator. In the asynchronous system, the base station retransmits the change in the state of the contention channel so that the mobile units can operate on this information. In a simple slotted system the base station also transmits slot timing. In a more complicated slotted system the base station also reserves slots for voice sources, detects and reports collisions, and provides priorities in the conflict resolution algorithm. The operation of a base station in a slotted system is more complicated than in the asynchronous system. In the appendix pseudocode is used to specify the operation of the base station. In this implementation, the first packet of a voice connection is distinguished from a data packet by the redundancy check that is used for the first minislot of the packet.

6. Conclusion

The performance of channel contention algorithms that are based upon observing the channel improves as the size of cells decreases. MSTDM makes it possible to take advantage of the channel efficiency of these algorithms while preserving the quality of voice conversations.

MSTDM should not be used in cellular radio networks in the same way that it was used in wired LAN's. Significant improvements are obtained by adapting MSTDM to the specific requirements of cellular radio networks:

By having the mobile units communicate with a base station, rather than directly with one another the transmitter power is reduced and the separation between cells that can use the same frequency is also reduced.

The bandwidth that is needed to implement the base station-centric system is cut almost in half because the primary objective of a mobile unit is to connect with the communication infrastructure and not with other mobile units in the same cell.

Finally, by using the base station for timing and control,

— the quality of service guarantees for voice are improved,

— the efficiency of data access is improved,

— the packet overhead is reduced, and

— the delay for the first packet in an active voice segment is reduced.

APPENDIX: *The Pseudocode for a Base Station in a Slotted System*

```
X_Idle()      Contention Channel is Idle
V_Rsv = Waiting if a voice reservation is waiting to be sent
Sets_Voice = number of voice sets in the contention tree
Sets_Data = number of data sets in the contention tree
    If(V_Rsv == Waiting) { X_V_Rsv() }
    If(Sets_Voice > 0) { Contention_Voice() }
    If(Sets_Data > 0) { Contention_Data() }
    Invitation_All()

X_V_Rsv()      Send out a voice reservation
    V_Rsv = Empty
    Xmit(Reserved)
    If(Channel==Busy) { R_list(Voice); X_Busy(); }
    Else { R_list(Empty); X_Idle(); }

Invitation_All()
    Xmit(Idle_Any)
    If(Channel==Idle) { R_list(Empty); X_Idle(); }
    Xmit(Busy)
    If(Channel==Idle) { R_list(Empty); R_list(Empty); X_Idle(); }
    If(Check_header_data==Good) { R_list(Empty);      R_list(Empty);
        X_Busy(); }
    If(Check_header_voice==Good) { R_list(Reserved);  R_list(Empty);
        X_Busy(); }
    Xmit(Collision)
    Sets_Voice = 1; Sets_Data = 1;
    R_list(Empty); R_list(Empty); R_list(Empty);
    If(Channel==Busy) {X_Busy()}
    If(Channel==Idle) {X_Idle()}

Contention_Voice()
    Xmit(Idle_Voice)
    If(Channel==Idle) { Sets_Voice--; R_list(Empty); X_Idle(); }
    Xmit(Busy)
    If(Channel==Idle) { R_list(Empty); R_list(Empty); X_Idle(); }
    If(Check_header_voice==Good) { Sets_Voice--;   R_list(Reserved);
        R_list(Empty); X_Busy(); }
    Xmit(Collision)
    Sets_Voice++;
    R_list(Empty); R_list(Empty); R_list(Empty);
    If(Channel==Busy) {X_Busy()}
    If(Channel==Idle) {X_Idle()}
```

Contention_Data()

Interrupt every k trys to give voice packets that have joined priority

```
If(trys++==k) { trys=0; Sets_Voice=1;}
Xmit(Idle_Any)
If(Channel==Idle) { Sets_Data++; R_list(Empty); X_Idle(); }
Xmit(Busy)
If(Channel==Idle) { R_list(Empty); R_list(Empty); X_Idle(); }
If(Check_header_data==Good) { Sets_Data--; R_list(Empty);
    R_list(Empty); X_Busy(); }
If(Check_header_voice==Good) { Sets_Data--; R_list(Reserved);
    R_list(Empty); X_Busy(); }
Xmit(Collision)
Sets_Data++;
R_list(Empty); R_list(Empty); R_list(Empty);
If(Channel==Busy) {X_Busy()}
If(Channel==Idle) {X_Idle()}
```

X_Busy() *Contention Channel is Busy*

```
Xmit(Busy)
R_list(Empty)
If(Channel==Idle) {
    If(Check_packet==Good) {Send an ACK to the Source}
    X_Idle() }
X_Busy()
```

R_list(Current) *Maintain the list of reservations*

R_table[n_table] Reserve a minislot n_table minislots after the beginning of a voice packet

i_table = the current entry in R_table

```
R_table[i_table]=Current; i_table++ mod n_table;
If(R_table[i_table]==Reserved) { V_Rsv=Waiting;
    R_table[i_table]=Empty; }
return
```

Xmit(Out)

```
Send "Out" on the outbound channel at the beginning of each
    minislot
    Out = Reserved, Collision, Busy, Idle_Voice, or Idle_Any
Set "Channel" to the state of the contention channel at the end of
    the minislot
    Channel = busy or idle
Set "Check_header_data" after the first minislot is received, based
    on the data header redundancy check
Set "Check_header_voice" after the first minislot is received, based
    on the voice header redundancy check
Set "Check_packet" after the entire packet is received, based on the
    packet redundancy check
    Checks = Good or Bad
```

REFERENCES

[1] N. F. Maxemchuk, "A Variation on CSMA/CD That Yields Movable TDM Slots in Integrated Voice/Data Local Networks," BSTJ, Vol. 61, No. 7, Sept. 82, pp 1527-1550.

[2] N. F. Maxemchuk, "Some Characteristics of Movable Slot TDM," Proc. 8th Conf. on Local Computer Networks, Minneapolis, Minn., Oct. 1983.

[3] F. A. Tobagi, V. B. Hunt, "Performance Analysis of Carrier Sense Multiple Access with Collision Detection," Computer Networks, Vol. 19, no. 7, July 1976.

[4] R. M. Metcalf, D. R. Boggs, "Ethernet: Distributed Packet Switching for Local Computer Networks," Commun. of the ACM, 19, No 7, July 1976, pp. 395-404.

[5] N. F. Maxemchuk, "The Use of Packets in Cellular Networks," TM HA6172000-971017-03

[6] L. Kleinrock, F. A. Tobagi, "Packet Switching in Radio Channels: Part I-Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics," IEEE Trans. on Commun., vol. COM-23, no.12, Dec. 1975, pp 1400-1416.

[7] F. A. Tobagi, L. Kleinrock, "Packet Switching in Radio Channels: Part IV-Stability Considerations and Dynamic Control in Carrier Sense Multiple Access," IEEE Trans. on Commun., vol. COM-25, no. 10, Oct. 1977, pp. 1103-1119.

[8] J. L. Sobrinho, A. S. Krishnakumar, "Distributed Multiple Access Procedures to Provide Voice Communications Over IEEE 802.11 wireless networks," Proceedings of IEEE Globcom, Dec. 1996, pp. 1689-1694.

[9] N. F. Maxemchuk, A. N. Netravali, "A Multifrequency Multiaccess System for Local Access," Proc ICC ’83, Boston, Mass., June 20-22.

[10] N. F. Maxemchuk, A. N. Netravali, "Voice and Data on a CATV Network," IEEE J. on Sel. Areas in Commun., vol SAC-3, No. 2, Mar. 1985 pp. 300-311.

[11] P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations," BSTJ, Jan. 1968, pp. 73-91.

[12] K. Bullington, I. M. Fraser, "Engineering Aspects of TASI," BSTJ, Vol. XXXVIII, No. 2, March 1959, pp 353-364.

[13] C. J. Weinstein, E. M. Hofsetter, "The Tradeoff Between Delay and TASI Advantage in a Packetized Speech Multiplexer," IEEE Trans. on Commun., vol. COM-27, no. 11, Nov. 1979, pp. 1716-1720.

[14] N. Abramson, "The Aloha System - Another Alternative for Computer Communications," Fall Joint Computer Conference, AFIPS Conference Proceedings, Vol. 37, pp. 281-285, 1970.

[15] L. Kleinrock, S. S. Lam, "Packet Switching in a Multiaccess Broadcast Channel: Performance Evaluation," IEEE Trans. on Commun, vol COM-23, no. 4, Apr. 1975, pp. 410-423.

[16] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard, B. Ramamurthi, "Packet Reservation Multiple Access for Local Wireless Communications," IEEE Transactions on Communications, Vol. COM-37, No. 8, pp. 885-890, August 1989.

[17] J. F. Hayes, "An Adaptive Technique for Local Distribution," IEEE Trans on Commun., COM-26, pp. 1178-1186, Aug. 1978.

[18] J. Capetanakis, "Tree Algorithms for Packet Broadcast Channels," IEEE Trans on Inform. Th., IT-25, 505-515, Sept. 1979.

[19] D. Bertsekas, R. Gallager, **Data Networks**, Prentice-Hall Inc, 1992.