Applying Packet Techniques to Cellular Radio

N. F. Maxemchuk

AT&T Labs - Research

ABSTRACT

The cell size in mobile networks is decreasing to accommodate more users in the same bandwidth. Smaller cell sizes increase the number of handoffs between cells and the probability of encountering a cell with more users than the bandwidth can accommodate.

Packet techniques can reduce the work that is performed in a handoff and reduce the probability of losing a connection in an over populated cell. In order to obtain these advantages, the mobile network must consist of both a wired and wireless segment and different packet techniques must be used for the inbound and outbound traffic on each of these segments. A fifth packet technique is used for a control channel that is independent of a mobile unit's location.

The resultant packet network has unique characteristics and requirements. We will show how these characteristics affect the structure of the wired network and the access protocols that are used in the wireless network.

1. Introduction

The number of users in mobile networks is increasing. Smaller cell sizes provide a means of supporting a larger number of users with the same bandwidth. As cell sizes shrink, users move between cells more frequently. In a circuit switched system each move requires that one circuit be torn down and another set up. Therefore, as cell sizes decrease the work associated with handing off users between cells increases. In addition, as a user traverses more cells he is more likely to encounter an overcrowded cell and lose the connection.

The "stateless" characteristics of some packet networks can reduce the work required for handoffs relative to "stateful", circuit switched networks. Packet addresses are used to reduce the work associated with changing cells by routing individual packets rather than setting up and tearing down circuits.

The ability of packet switching to spread information over the network and give priority to segments of the same stream makes it possible to degrade service, rather than denying service, in oversubscribed cells. When a cell is overloaded, some packets use adjacent cells or the packet rate is systematically reduced, rather than disconnecting sources.

In section 2 we describe the complete architecture of the cellular network. In order for a mobile network to address handoffs and loss of service between cells the network must include a portion of the network that interconnects the cells. In our architecture, the mobile network consists of a wired metropolitan area network (MAN) as well as the wireless transmission in the cell. The wired MAN must not only be designed to accommodate mobility between cells, but must also take into account the unique traffic patterns that occur in the cellular architecture, and the difference in the cost of bandwidth between a MAN and either the wide area network (WAN) or wired network. Several MAN technologies are applied to the wired portion of the mobile network in order to show the differences between the cellular network and earlier MAN's.

In section 3 we identify five networking problems associated with the mobile network architecture. Each of the problems requires the applications of a different packet technology. The architecture consists of a wired and wireless segment and each segment provides outbound communications to the mobile unit and inbound communications from the mobile unit. Four different packet techniques are used in these segments. The fifth packet technique is used for a universal control channel to place calls, notify the mobile units of cell boundaries and redistribute the load during overload conditions.

One of the earliest applications of packet technology was in wireless networks[1,2]. Packet techniques are used in military radio networks, where survivability is a principal concern, and in wireless LAN's[3], that are primarily designed to interconnect data terminals in a limited area. Packet techniques are not generally used in commercial cellular networks. The cell sizes are too large to use the same contention mechanisms as in LAN's, and the quality of a voice connection is more important than the efficiency of data transmission.

As cell sizes become smaller, packet contention mechanisms become a viable means of sharing the cellular radio channel. There are, however, significant differences between cellular networks and either the LAN or military applications.

1. Cells in a cellular network have a basestation that can provide timing, coordination and signal regeneration. Military networks in particular avoid a single point of failure.

2. The primary objective in a cellular network is to connect wireless units with the communication infrastructure rather than with nearby wireless units. Wireless LAN's in particular are designed to interconnect the units within a local area.

These differences result in significant changes in the access protocols. All of the protocols that we consider are variants on the MSTDM[4] protocol and provide the necessary service guarantees for telephone quality voice connections. Because of the differences between cellular networks and the earlier applications, up to four time as much bandwidth is available in each cell and there can be up to eight times as many voice connections. The protocol variants are outlined in section 4. There is a more detailed description of the variants in reference 5.

In section 5 we deal with overcrowded cells. In a channelized network a voice connection that enters an oversubscribed cell is dropped. In our packet network there is both data and voice traffic. As a cell becomes overcrowded, the channel sharing techniques reduce the the bandwidth available to individual data terminals, as in the Ethernet and most packet protocols that contend for a channel. In order to reduce the probability of disconnecting a voice terminal in an overcrowded cell, three levels of oversubscription are described and the actions that are taken depend upon the level. Voice streams are split into multiple packet streams that define different quality of services and the streams are treated differently.

In section 6 we look at the performance of the network as we vary a number of parameters. In addition to the conventional measures of voice and data throughput, we consider the effects of the stream splitting techniques that are introduced to deal with oversubscription.

The resulting packet system significantly reduces the work associated with handoffs, and is much less likely to experience a service disruption caused by an overload than current cellular networks.

2. Architecture

Figure 1 is the architecture of the cellular network. The radio network consists of cells, the hexagons, that completely cover a service area. The service area can span any convenient geography. For instance, the service area may span a city, a city and its suburbs, or an area in which people frequently commute. In each cell there is a basestation that provides the connection between the radio network and a wired network. The wire lines in a service area connect the basestations and the switching center (SX). The switching center connects the cellular network with the existing communications infrastructure, the WAN. A service area may include several switching centers.

Each active mobile unit is represented by an agent at the interface to the WAN. The agent stays in the same location, as long as the mobile unit is in the service area, and makes it unnecessary for the WAN to respond to movement between cells. The WAN must still respond when a mobile unit changes service areas, and the agent is moved, but this should be a relatively infrequent event and is not affected by decreasing the cell

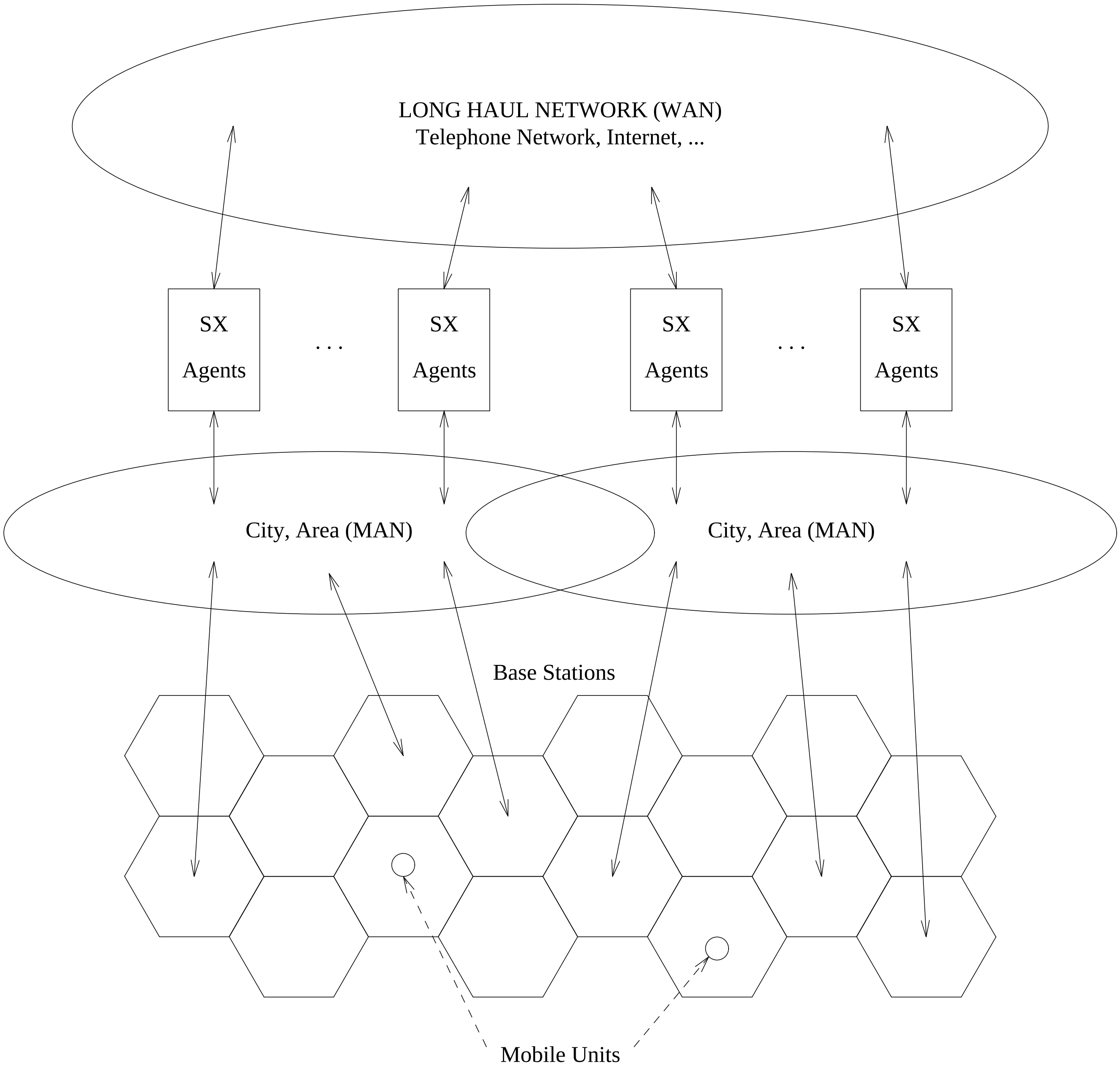

Figure 1. Network Architecture

size within a service area.

An agent maintains a connection on the WAN on behalf of the mobile unit. The agent breaks the connection at the end of a communication session or when a failure occurs. Since the agent is not always notified of a failure, it maintains a soft state connection. If the agent stops receiving packets for a period of time it first trys to contact the mobile unit. If communication with the mobile unit cannot be reestablished, the connection on the WAN is terminated. Resources in the WAN are not tied up indefinitely.

The agent is responsible for translating between the data and voice formats that are used in the switched and mobile networks and for all operations needed for mobility within the service area. For instance, the agent resequences the packets received from the mobile unit, if necessary, and retimes the voice samples before converting them to the coding used on the WAN. The agent also recodes the voice samples from the WAN, puts them into packets, and addresses the packets to the proper base station.

A basestation sends the packets from agents over the radio channel and sends packets received on the radio channel to an agent. The basestations do not switch packets to other basestations for communications between mobile units in the service area. Therefore, for each mobile unit the destination is at a fixed location and the basestation does not track mobile units. Routing all communications through the agent simplifies the switching functions performed by a basestation. The decision to communicate with a fixed agent is justified because the primary objective of the network is to connect a mobile unit to the communications infrastructure.

The mobile units only send packets to, and receive packets from, the basestation. They do not receive packets directly from the other mobile units that share the radio spectrum of the cell. If two mobile units that are communicating happen to be in the same cell the communication path flows through the basestation, to an agent, through the WAN (a local call), to another agent, back to the same base station and over the radio channel. This is a much longer path than the direct path over the radio channel, however, communications between two mobile units in the same cell is not the intended use of the network, as in a LAN. Using the longer path for a small fraction of the connections makes more efficient channel access techniques available to all mobile units, as shown in section 4.

2.1 The Wired MAN

A wired MAN connects the agents and basestations in our cellular network. This network is not considered part of many wireless networks, however, it must be part of any network that removes responsibility for mobility from the WAN. The requirements on the MAN include the traffic distribution, reliability considerations, load redistribution, and a need to limit the migration of agents.

The traffic on the MAN is between basestations that are distributed over the service areas and a small number of centers where the agents connect to the WAN. All of the traffic from the service area is funneled into the centers and all of the traffic from the centers is spread out across the service area. There is no traffic between basestations. The traffic pattern is similar to that between the telephones and switching center in the telephone network or between the customers and edge switches in an ISP's network. The pattern is very different from that on data MAN's. Data MAN's, such as the FDDI[6] dual ring network, the DQDB[7] dual bus network, and the Manhattan Street Network[8](MSN), are designed to provide communications between all of the stations on the network. Most data MAN's perform best with uniformly distributed loads, or traffic within communities of users.

The endpoints on the cellular MAN are basestations that connect to a large number of users, rather than the single user in the telephone network or an ISP's network. A basestation should be multiply connected to the MAN, as is a large PBX, to provide an appropriate level of reliability.

Multiply connected networks can redistribute the load to avoid heavily used paths. If there is one path from the switching center to each basestation, the path must carry the peak load. When there are multiple paths and an ability to redistribute the load, the paths can be designed to carry less than the peak load. As a result, it is possible to design more economical MAN's.

When a service area has several switching centers that connect with the WAN there will be adjacent cells that are served by different switching centers. When an agent moves between switching centers, the connection in the WAN must change. The topology of the MAN must make it unnecessary for an agent to migrate as soon as the mobile unit crosses a cell boundary. Otherwise, a mobile unit that oscillates back and forth across the cell boundary will have its agent oscillate back and forth between switching centers. There should be capacity on the MAN that connects the switching centers so that an agent is not moved until the mobile unit is well into the area that is supported by a new switching center.

The combination of requirements are significantly different from those on previous MAN's. Some of the requirements are similar to those on the local distribution networks that have been installed by the telephone and CATV industry, and others are similar to those on data MAN's. Rather than designing a new network, we will show how existing networks meet or fail to meet the requirements and how existing networks can be combined to meet all of the requirements. There are a number of ways that the networks can be combined and the characteristics of the resulting network provides a potential means of distinguishing competing cellular systems.

The tree topology used in CATV networks can support the traffic distribution in the wired MAN, and provide line sharing. In the tree topology the head end is located at the WAN switch and the cells at the branches. At each split in the tree, routing decisions are made for packets destined for the basestations and multiplexers combine the packets directed toward the central office. At the central office a demultiplexer splits the traffic for each agent. An advantage of this approach is that is that the bandwidth can be tailored to the non-uniform load of this application by increasing the bandwidth near the head end of the tree. Store-and-forward techniques that are commonly used in packet networks can be used at the multiplexers and demultiplexers to share the bandwidth among the basestations, or packet techniques can be used that are specifically designed for CATV networks[9]. The principle disadvantage of this architecture is the reliability. There are a many locations in the tree where the failure of single line or component can disrupt communications for a large number of cells.

The current telephone network is a star topology. This network can become the wired MAN by using digital subscriber loop (DSL) technology to multiplex the packets to basestations onto a small number of lines. The lines from each basestation terminate at a packet switch. The packet switch performs a multiplexing and demultiplexing function that is similar to the edge switches that are currently used by the ISP's. The agents are between the packet switch and the circuit switch on the WAN. As a mobile unit moves between cells the connection through the packet switch changes but the connection on the WAN remains fixed. The basestations in the star network do not share bandwidth. Each line from the basestation to the central switch must accommodate the peak bit rate from the basestation. There is a single path between the basestations and the central office, which affects reliability. Hysteresis can be introduced into this network by interconnecting the packet switches at adjacent switching offices.

The reliability and bandwidth sharing in the star network is increased by inserting a layer of routers between the edge switches that are connected to the agents and the edge switches that connect to the basestations, just as in the Internet. The routers are multiply connected with wide bandwidth pipes. Wide bandwidth pipes can be provided over SONET rings and are therefore more reliable than the lower bandwidth connections to individual basestations. The pipes between routers are shared by many basestations and the load can be redistributed to reflect changes in load from the basestations. Routers are too expensive to locate one at each of a large number of small cells. Therefore, the final connection to the basestations must still be an edge switch. The network reliability can be improved by connecting each basestation to several edge switches at different routers.

The reliability and sharing for the final connection to the basestation can also be improved by using any of the three MAN technologies that have been mentioned. The MAN technologies may be more economical than multiple edge switches, while improving sharing and reliability. Both DQDB and FDDI networks can survive single failures. However, the number of basestations that can share a particular DQDB or FDDI network is constrained because the throughput per basestations decreases linearly with the number of basestations. The DQDB network can be partitioned to reuse the bandwidth, however, bandwidth re-use works best with traffic that remains in a community of interest[10], and will not provide a significant advantage with our particular non-uniform load.

The MSN is inherently more reliable than DQDB and FDDI[11]. When the sources and destinations are uniformly distributed over the MSN, the capacity of the MSN decreases as the square root of the number of nodes, instead of linearly. In addition, when the sources and destination are located near one another, the MSN is partitioned into non-interfering communities of interest, which makes it possible to support an arbitrarily large numbers of nodes. Agents can appear to be in the same neighborhood as their basestation, even the as the mobile units move, by connecting the routers to the MSN in each neighborhood, as shown in figure 2, and sending packets for the agent to the router that is connected to the neighborhood of the basestation.

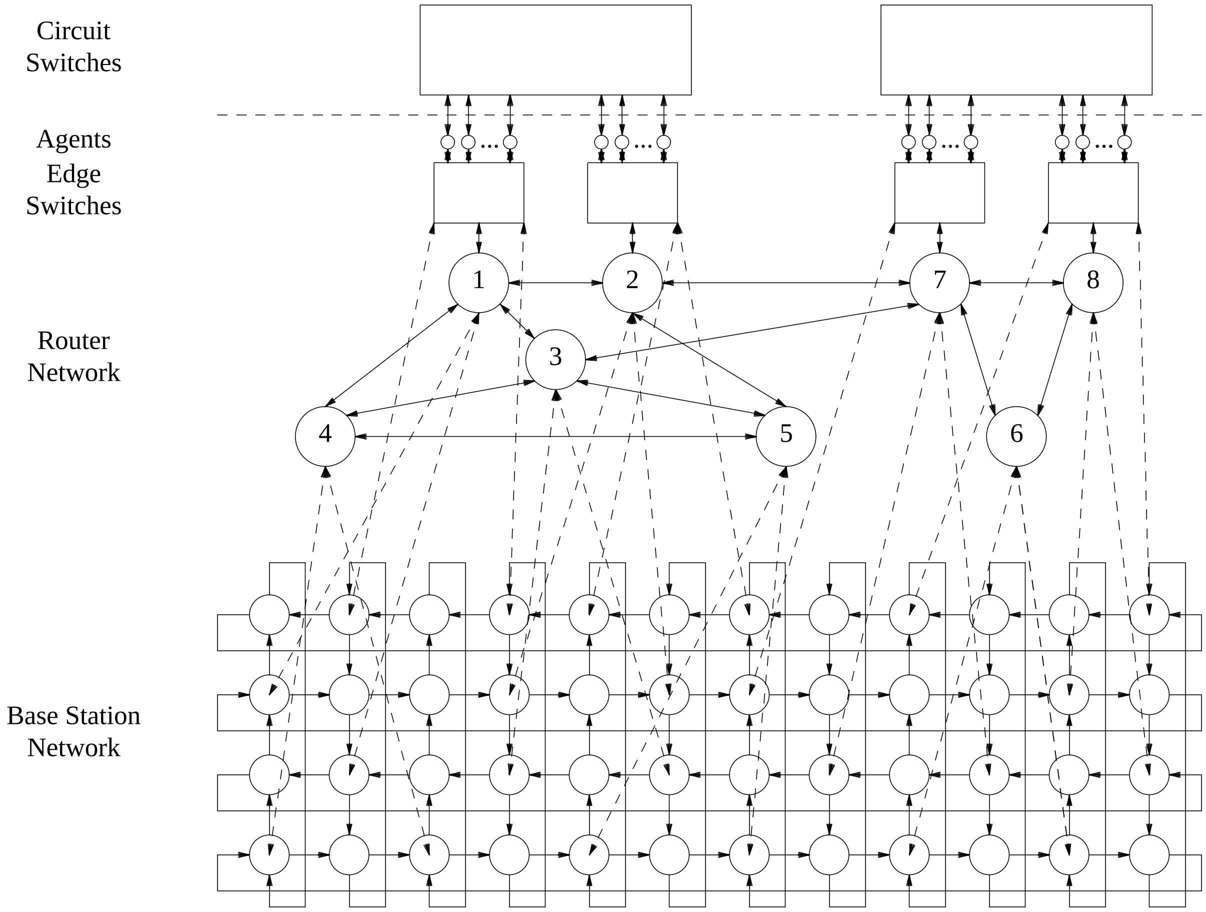

Figure 2. A Distributed Metropolitan Area Network

3. Mobility Control

A service area consists of a wired MAN and a wireless cellular system. Communications in these two portions of the network is separated into an inbound channel from the mobile unit to the agent and an outbound channel from the agent to the mobile unit. Different packet techniques are applied to the four channels. In addition, there is a control channel, that is shared by all of the base stations, that uses a fifth packet technique.

As the mobile user moves between the cells in the service area the path in the WAN remains the same. The work involved in a handoff between cells is reduced from tearing down and setting up a circuit that reserves resources, to changing the address on a packet tunnel over one quarter of the path. Packet techniques eliminate all of the work involved in a handoff on the remaining three quarters of the path.

3.1 Wireless Network

The wireless portion of the network consists of cells that cover the geography of the service area. A basestation that transfers packets between the wired and wireless portion of the network is located in each cell. The basestation is responsible for relaying packets between the mobile unit and its agent.

The frequency spectrum is divided into channels that are assigned to cells. The set of cells that are assigned the same channels are selected so that they do not interfere with one another. For the purpose of this discussion, all of the cells are the same size and shape and the basestation is located in the center of the cell.

In current systems, cells have irregular shapes that are determined by physical objects that limit signal propagation, and, the location of the basestation is determined by community objections. The main reason that we assume regular cells is to create a tractable problem that can be used as a basis for comparison. However, as cell sizes become smaller, there are fewer physical obstructions in a cell, which makes the cells more regular, and the signal power and antennae size decreases, which makes it easier to locate basestations near the center of the cell.

A cell has inbound channels, outbound channels and a control channel. The inbound channels provide communications from multiple mobile units to a single basestation. Multiple access packet techniques are used to share an inbound channel among geographically distributed users. The outbound channels are used to provide communications from a single base station to multiple mobile units. The basestation multiplexes the packets for multiple units on a single channel and packet addresses are used to identify the destination. The control channel is time shared by the basestations, and the packet address identifies the basestation that is transmitting.

3.1.1 Common Control Channel: The control channel provides unidirectional communications from the base stations to the mobile units. It is a single channel that the mobile units can receive anywhere in the service area. The channel is used to identify the closest basestation and the frequencies that it uses for the inbound and outbound channels. The channel is also used to locate mobile units when a call is placed. An agent locates an inactive mobile unit, to place a phone call, or an active unit, that has lost contact, by broadcasting a message to all of the basestations, each of which transmits the request on the common control channel.

All of the basestations transmit in the same frequency band, without interfering, by transmitting their packets during different time slots. Many basestations transmit in the same time slot. However, basestations that transmit in the same slot are far enough apart that they do not interfere with one another.

Figure 3 shows a pattern for reusing time slots. Time is divided into three slots and all of the basestations in cells with the same number transmit in the same time slot. For example, the basestation in the center cell transmits in all of the "1" slots. It must transmit enough power to be received by mobile units in the smaller circle, but not so much power that it interferes with the signals received by mobile units outside the larger circle. The mobile units that are outside the center cell and inside the larger circle receive control signals from basestations that use time slots "2" or "3". The radius of the outer circle is twice the radius of the inner circle, which provides a guard band on the transmit distance of 100%. The same radius circles can be drawn around any cell in the system.

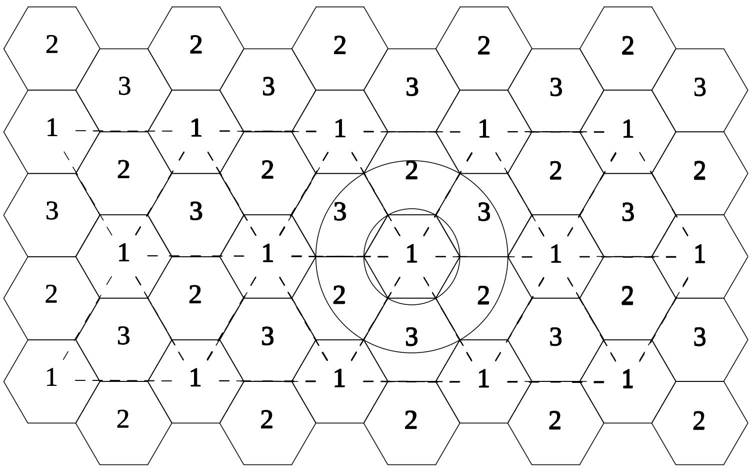

Figure 3. Time Slot Reuse of the Control Channel

3.1.2 In-bound traffic: All of the mobile users in a cell share the inbound channel to communicate with the basestation. Systems that reserve a frequency channel or time slot[12] for a user must allocate and deallocate channels as mobile units enter and leave the cell. Tracking the frequently changing channel assignments and determining when mobile units have left a cell is part of the handoff problem. Packet techniques circumvent the problem by providing resources on demand, rather than assigning a channel.

As the cell size decreases, contention based protocols that examine the channel state before and during transmission provide a means of sharing the channel without setting up and tearing down connections. The specific characteristics of cellular networks make it possible to implement more efficient contention based techniques than in other radio networks. In section 4 we describe the characteristics of cellular networks and the resulting modifications of existing protocols. The recommended protocols provide quality of service guarantees for voice.

In conventional cellular systems a connection is dropped when a mobile unit moves into a cell that does not have any available channels. In a packet system it is possible to decrease the quality of service rather than losing the connection. During heavy loads, a basestation can instruct all, or specific, mobile units to reduce their packet rate so that additional mobile units can transmit. It is also possible to redistribute the load in a cell by reassigning some packets or all of the packets from a mobile unit to other basestations. Specific methods for splitting the packet stream from individual mobile units to survive oversubscribed cells are described in section 5.

3.1.3 Out bound traffic: This channel is used to transmit from the single basestation in a cell to many mobile units. This is a much simpler problem than communicating in the other direction because there is a single source that multiplexes the packets onto the channel. The basestation forwards addressed packets from the agents to the mobile units. The mobile units listen to the channel for their own address and filter out the rest of the packets. There are no resources that are allocated or deallocated as mobile units change cells.

The basestation transmits a signal on the out bound channel, even when there is no information. Transmission is very efficient because there are no guard bands or timing acquisition intervals between the packets that are destined for different receivers.

There is a privacy problem. Mobile units can eavesdrop by receiving packets with other addresses. This problem is similar to that in current systems in which units can listen to channels other than their assigned channel. An advantage of the packet system is that the data is digital and can be encrypted or scrambled more easily than analog signals.

3.2 Wired Network

The wired MAN is a statistically multiplexed, packet switched network. The network has a variable delay and a nonzero probability of loss. The lower the utilization of the packet switched network, the higher the probability that a packet will be received before the deadline when it is scheduled to be used, and the better the quality of a voice connections. We assume that there is sufficient bandwidth in the wired MAN to obtain a reasonable quality of service without maintaining circuits or reserving resources. The assumption is based upon the fact that wide bandwidth circuits are more likely to be available, and are less expensive in this portion of the network than in either the cellular network or the long haul network on either side. Therefore, this portion of the network should not be the bottleneck.

Packets on the MAN are assigned different priorities for data and different voice segments. Voice has priority over data because data can tolerate variable delays better than voice. A voice stream is partitioned into high and low priority packets to survive oversubscription in a cell, as described in section 5. If portions of the wired MAN are congested lower priority voice packets are discarded.

3.2.1 In-bound traffic: Traffic from the mobile unit to its agent enters the wireless network at different basestations as the mobile unit moves between cells. Each packet has the address of the agent. Getting the packets to the proper agent, independent of the basestation where it enters the network, is a straightforward application of packet routers. It is not necessary to establish circuits or perform handoffs when mobile units change basestations.

With packet switching, packets can arrive out of order and interpacket timing is not be maintained. Packets must contain a sequence number and timing information so that the agent can accurately reconstruct the signal before transmitting it on the switched network. The RTP[13] protocol, used on the Internet, includes the necessary information.

Packet misordering and delay variations are more pronounced when a mobile unit changes basestations and transmits through a different part of the wired network. In section 5 we describe techniques that reduce the effects of cell boundaries in the wired network by delaying handoffs in the wireless network until a silent interval in the speech.

3.2.2 Out-bound traffic An agent keeps track of the basestation that can transmit to its mobile unit. The packets from an agent are addressed to a basestation and encapsulate the packets that the basestation transmits to the mobile unit. As a mobile unit moves from cell to cell it notifies its agent, and the agent changes the address on the encapsulation packet.

There are no network resources dedicated from the agent to the basestation. The only work involved in a handoff, when a mobile unit changes cells, is changing the destination address on the tunnel from the agent to the basestation.

In this approach, mobility is handled at the edge of the wired network. The routers transfer packets from an agent to a basestation, each with a fixed network address. This differs from many mobile IP[14] approaches which handle mobility inside the network.

4. Random Access Strategies

In this section we use random access strategies to share the inbound wireless channel among the mobile units in a cell. The recommended strategies provide quality of service guarantees for voice while sharing the channel with bursty data sources. The particular characteristics of cellular networks make it possible to modify previously used strategies to carry up to eight times as much traffic as on other radio networks. The primary objective is to show the differences between cellular radio and other radio networks. A more complete description of the recommended access methods, including the pseudocode to implement several of them, is in reference 5.

Many random access techniques have been used to share radio channels. The first were slotted and unslotted Aloha[1], which were improved using capture mechanisms[2,15]. This class of random access technique makes inefficient use of the channel.

A difference between our current cellular system and the early applications is the cell size. We assume that the cell sizes are becoming smaller to accommodate a larger number of users in the same bandwidth. When the propagation delay is small relative to the time it takes to transmit a packet, we can use more efficient strategies that detect the channel state, such as carrier sense multiple access[16](CSMA) or CSMA with collision avoidance(CSMA/CA),[3] that is the IEEE 802.11 standard for wireless LAN's. The primary applications of CSMA and CSMA/CA have been to interconnect nearby data terminals. The terminals in these networks communicate directly with one another over the radio channel.

By contrast,

— cellular networks are designed to connect mobile units with the communications infrastructure with little or no direct communications between mobile units in a cell,

— each cell has a basestation that can be an intermediary or mediate communications, and

— the network must provide good quality voice communications in addition to data.

The basestation makes it possible to implement CSMA with collision detection(CSMA/CD)[17] that is used on the Ethernet[17]. Collision detection is much more difficult to implement in a radio network than in a terminated, cable network. In the radio network reflections of a station's signal can be much stronger than the signal from other stations. Compensating for reflections in a mobile system requires considerable processing and delay that is incompatible with the strategy of quickly transmitting a packet then releasing the channel.

In the cellular network the basestation retransmits the signal that it receives on the contention channel in a separate frequency band. The mobile units detect collisions by listening to the second channel. The signal on this channel provides more information than conventional collision detection. Since the basestation is the intended receiver, the signal on this channel indicates that the signal was correctly received.

The two channel version of CSMA/CD has been used in CATV networks.[9] in which the directional transmission characteristics prevent stations from directly receiving the signal from all contending stations. In order to implement CSMA/CD, stations transmit upstream in a contention channel and the head-end retransmits the signal downstream on a reflection channel that all of the stations receive.

Two channel CSMA/CD uses twice the frequency spectrum as single channel CSMA. However, the signals do not have to be transmitted as far and the distance between cells that can reuse a frequency band is smaller. In section 4.1 the relationship of the area in a cell to the area between cells that use the same frequency is developed. The ratio of the useful area to the total area is four times higher in the two channel system than the single channel system. The bandwidth required per channel is doubled while the channels per area may be quadrupled. As a result, twice as many channels are available in an area.

The factor of two improvement is obtained under idealized conditions that are necessary for analysis, but seldom approached in a real system. In section 4.2 we discuss factors that effect the actual improvement. Some of the factors, due to a single receiver, may increase the factor of two. Other factors, such as developing a frequency reuse plan may increase or decrease the factor of two.

Collision detection makes it possible to implement a variant of CSMA/CD, movable slot TDM (MSTDM)[4], that provides good quality of service guarantees for voice communications. In MSTDM:

— Data sources and the first packet in a stream of voice packets use the standard CSMA/CD protocol.

— A continuation voice packet is transmitted a fixed period, T_V seconds, after the successful transmission of the previous packet and only uses CSMA.

— If the channel is busy when a continuation packet is ready, the packet is transmitted as soon as the channel becomes idle and Samples that arrive while waiting are transmitted in an overflow area. The source does not need to acquire the channel for T_V seconds after either an immediate or delayed transmission.

— A continuation packet has a preempt signal at the beginning of the packet so that a data source that collides with it can detect the collision and stop transmitting, before it interferes with the voice.

By constraining the size of data packets, voice sources never collide with each other, even when the utilization approaches one[4]. Therefore, there is no distortion of the voice source and the only voice delay is the packet assembly time.

A second factor of almost two improvement is obtained because the primary objective of a cellular system is to communicate from the mobile units to the basestation and not between mobile units in the cell. Therefore, the reflection channel is only used to determine the state of the contention channel, and not to carry data. In section 4.3 we show a variation on MSTDM that operates with three channel states and describe a means of multiplexing the channel state onto the outbound channel. Therefore, we almost completely recover the reflection channel.

The small amount of state information used by the variant of MSTDM causes a slight decrease in the efficiency of data transmission, because data packet collisions are not detected, and a slight degradation in the quality of service for voice, because there is an infrequent event that causes voice packets to collide. Both decreases are recovered by the minislotted system introduced in section 4.4. Slotted system are easy to implement in cellular networks because the base station transmits continuously on the outbound channel and provides timing. In addition to recovering the losses that are caused by the reduced state information, the minislotted system recovers the preempt headers that were introduced in MSTDM and provides faster contention resolution.

Both the asynchronous and slotted variants of MSTDM provide a means of assigning priorities to the sources that contend for the capacity that is left after the voice sources take their share. In both systems the packets that establish a position for new voice sources are given priority over data packets. As a result, voice sources can release the channel during a silent interval and quickly reacquire the channel at the beginning of the next active interval. Removing the silent intervals results in the final factor of two improvement[18], the TASI[19] advantage.

In order to maintain speech quality, TASI systems constrain the number of voice connections so that the channel is reacquired without clipping 95% of the time. An equivalent constraint on MSTDM is that 95% of time the first packet in an active voice sequence must acquire the channel in less than a packet assembly time. This constraint determines the maximum load on the channel. For instance, if the time to assemble a packet is 20 milliseconds and we use the Ethernet retry strategy, with retrys separated by more than 20 milliseconds, 95% of the initial voice packets must acquire the channel on the first try. In order to maintain TASI quality, the channel utilization must be less than 5%. Clearly, with this retry strategy it does not make sense to suppress silent intervals. Our retry strategies quickly reestablish voice connections, even under heavy loads, and make it possible to increase the number of voice sources nearly to the TASI limit.

4.1 Frequency Reuse

Figure 4 shows the relationship between the area in a circle that uses a frequency band and the area that cannot reuse that band. In each figure the areas that use the same frequency bands are the small circles. A mobile unit may be anywhere within this circle and the basestation is located at its center. The distance between the circles is adjusted so that the signals in different circles do not interfere.

The centers of the small circles are arranged in an hexagonal array so that the distance between all neighboring circles is the same. This geometry provides the maximum packing. The line between the center of each circle is bisected to form the dashed hexagons. The area of the circle divided by the area of the hexagon is the fraction of the surface that can use a frequency band.

In the two channel system mobile units do not communicate directly. The mobile units transmit in a contention channel. The basestation receives the signal in the contention channel and retransmits it in a reflection channel. The mobile units receive the signal in the contention channel and use it to determine when the contention channel is busy and when the signals they transmit are successfully received by the basestation.

The signal in the reflection channel from a basestation must be received by all of the mobile units in the circle, but must not interfere with the signal that is received by a mobile unit in a different circle. In order to make certain that the signal from the proper basestation is stronger than that from other basestations, there is a guard band αr between circles of radius r. In figure 4 $\alpha = .3$.

The signal in the contention channel from the mobile units must be received by a basestation that is up to r

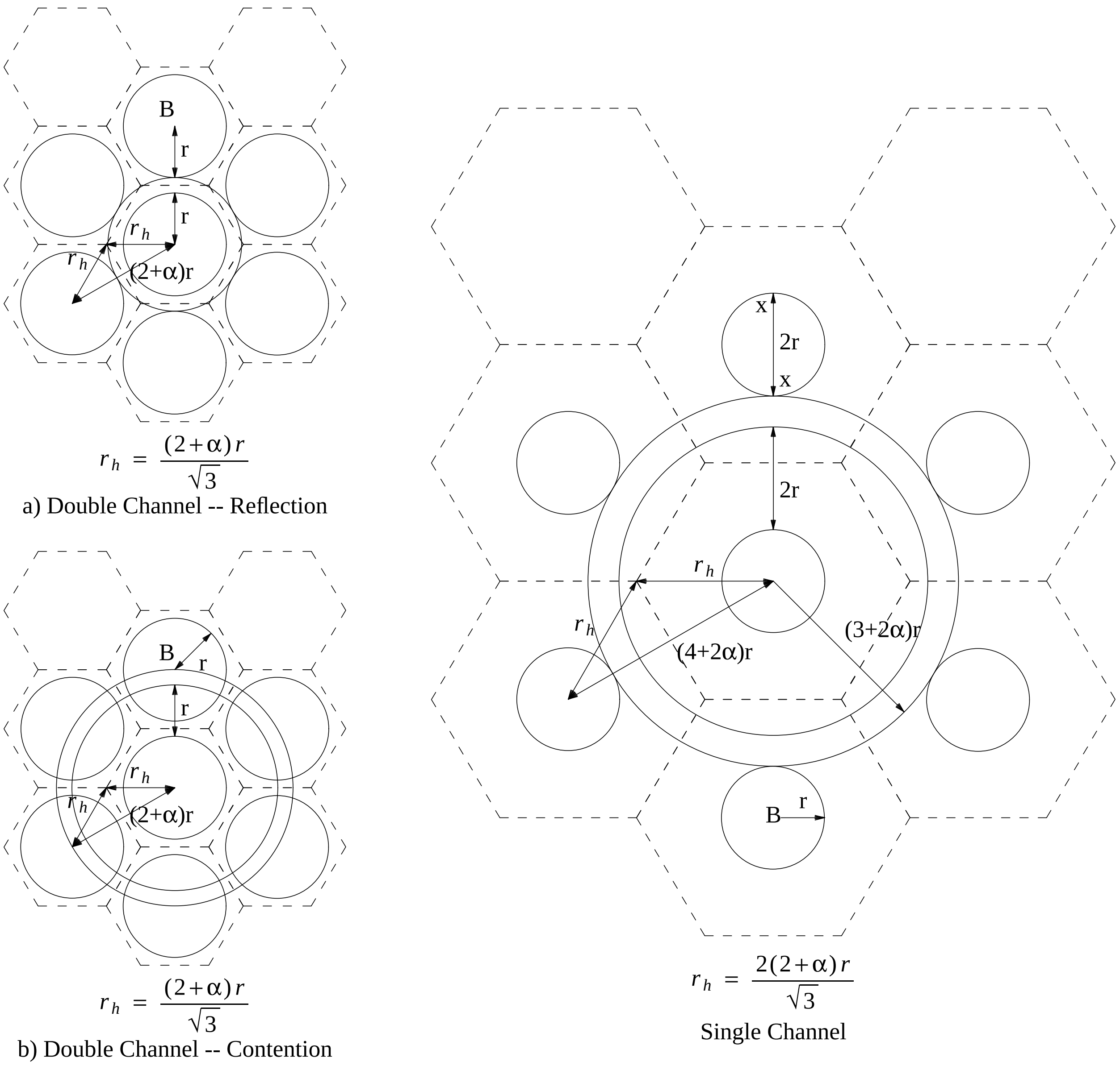

Figure 4. Frequency Reuse in Single and Double Channel CSMA

away, but must not interfere with the signals received at a basestation that is further than $(1+\alpha)r$ away. Looking at the figures, the separation between circles is the same for the contention and reflection channels. The ratio of the area of the circle to the area of the encompassing hexagons is $R_2 = \dfrac{2\pi}{\sqrt{3}\,(2+\alpha)^2}$.

In the single channel system mobile units in a circle must detect the signal from all of the other units in their circle. The signal from a mobile unit in an adjacent circle should be weak enough that it is not confused with the signal from a station in the same circle. The distance between the edges of adjacent circles is $2(r+\alpha)$.

The ratio of the areas in the single and double channel systems is $R_1 = 4 * R_2$, independent of the guard bands. In order to transmit at a particular bit rate, the two channel system uses twice the bandwidth of the single channel system, however, the two channel may reuse the bandwidth over four times as much of the service area. Therefore, if there is a fixed amount of bandwidth assigned to a service area, the two channel system can assign four times as much bandwidth to a transmission area as the single channel system and achieve twice the bit rate.

4.2 Reuse Considerations

The factor of two in the previous section is obtained under ideal conditions that were assumed for analytic simplicity. The factor may increase or decrease in more realistic systems. Practical considerations that effect the factor of two include:

1. the shape of the cell and the location of the basestation,
2. the amount of processing performed by mobile units and the amount of control exercised by the basestation, and
3. the frequency plan that reuses the bandwidth.

The cells in practical systems are irregular and the basestation may not be near the center of the cell. The shape and size of a cell reflects physical objects that interfere with transmission and usage patterns due to population centers, roadways, etc. The antennae associated with basestations are large, highly visible objects that are located where they are allowed, rather than in the best location in a cell. As cell sizes become smaller, there are fewer reasons to create irregular cells or to misplace a basestation. Usually, in a smaller areas there are fewer physical obstructions or nonuniform usage patterns. The basestations used in smaller cells have smaller antennae that are easier to position near the center of a cell. Therefore, systems with small cells are more likely to conform to the ideal conditions than current cellular systems.

Irregular cells can be constructed that improve or degrade the performance of two channel systems relative to single channel systems. As an example of an irregular structure that degrades the performance, locate the basestation on the circumference of the circle, instead of the center. A mobile unit in the two channel system must transmit enough power to reach across the diameter, rather than the radius, of the cell. The mobile unit in a two channel system must transmit as far as in the single channel system. The contention channel in the two channel system covers the same surface as in the two channel system, and the area covered by the reflection channel reduces the bandwidth available in a transmission area. Instead of the bit rate in the two channel being twice as large as in the single channel system, the bit rate is lower than in the single channel system.

Standard techniques reduce the negative impact of irregular cells on a system with an off center basestation. The techniques change the footprint of the signal from the basestation and the footprint of the signals from the mobile units to conform more closely to the shape of the irregular cell. In current systems, when the basestation is located off center or the cell is irregular in shape, a directional antennae is used to form the beam from the basestation to fit the the cell. In CDMA networks the near/far problem is solved by controlling the transmitter power from the mobile units so that the signal power that reaches the basestation from all units is approximately the same[20]. By using the same technique in an irregular cell, the footprint of the signals from the mobile units more closely matches the shape of the cell than if mobile units transmit with the maximum needed power independent of location.

In the future we expect the improvement in basestation systems to increase further by detecting the location of the basestation from its outbound channel and using an antenna array to shape the beam from the mobile units toward the basestation. This operation is not currently performed because of the power required by a processor that continuously calculates the multiplicative coefficients for the antennae array as the mobile unit moves and applys those coefficients to the arriving signals. As cell sizes become smaller, less power is required for the transmit signal, and more power is available for processing.

The third practical consideration is the packing problem. We cannot always configure cells that reuse the frequencies as often as they may be reused. The ability to pack the cells can change the relative improvement of the single and double channel systems. To explain the problem we return to the ideal situation. The circle that defines a transmission region is replaced by a hexagon with the same radius. The service area consists of hexagons that are all the same size. Our objective is to select the sets of hexagons that use the same frequencies. We must pick a set of hexagons that are as close together as possible, but satisfy the constraints on the distance between transmission circles.

In the two channel system, adjacent hexagons cannot use the same frequency, even when the guard band $\alpha = 0$ because the transmission circle extends beyond the edge of the hexagon on each side. The time slot reuse plan in figure 3 is the most compact plan for reusing frequencies in the two channel system. The frequency spectrum is divided into three parts and $(1/3)^{rd}$ of the frequency spectrum is used in each cell. The minimum distance from the center of a cell to another cell that uses the same frequency is r, the radius of the hexagon, while the actual distance is $2r$, therefore, the guard band, α, is 100%.

Figure 5 shows the most compact frequency plan that satisfies the distance constraints for a single channel system. The minimum distance from the center of a cell to the edge of another cell that uses the same frequency in a single channel system is $3r$ while the distance in this frequency plan is $3.66r$, therefore, α is 22%. The frequency spectrum is divided into 7 parts and $(1/7)^{th}$ of the spectrum is available in each cell. If the 22% guard band is adequate, then $(7/3)^{rd}$ as much bandwidth is available in a cell in the two channel system as in the single channel system, instead of 4 times as much bandwidth. If the guard band must be more than 22% but less than 100%, then the plan in figure 3 can be used for the double channel system, but a plan with a greater spacing will be required for the single channel case. The bandwidth for the single channel system will be divided into more than seven parts, and the ratio will be larger.

4.3 Asynchronous Protocols

The contention channel in a cellular system is used to communicate from the mobile units in a cell to the basestation. There is no direct communication between mobile units. Therefore, the reflection channel is only used to determine the state of the contention channel. In a CSMA system the information that is needed is a single bit that indicates when the channel is busy. The limited information on the reflection channel is multiplexed onto the broadcast channel, that is used to communicate from the basestation to all of the mobile units, instead of using a separate frequency band..

The state of the contention channel changes asynchronously with respect to the packets on the broadcast channel. In order to report the change in state quickly we interrupt the broadcast transmission with a data link escape sequence (DLE). A DLE is a unique character that indicates that the next character is a control character. One function in the control sequence indicates that the DLE occurred naturally in the data. There are both bit and byte oriented DLE protocols. The bit oriented protocols report changes more quickly, because they do not wait for a byte boundary, but the byte oriented protocols are preferable because they are currently less expensive to implement.

Much more information is needed in a CSMA/CD system than a CSMA system. In a CSMA/CD system the bits at the beginning of the packet are needed to distinguish a collision from one’s own transmission. As the cell size becomes smaller the throughput that can be obtained with CSMA and CSMA/CD become closer. The primary reason for implementing collision detection in our cellular network is to obtain the service guarantees of MSTDM for voice communications.

A variant of MSTDM has been found that requires almost as little contention information as CSMA.[5] The basestation only reports three channel states, idle, busy with data only, and busy with voice. The basestation distinguishes data and voice sources by the clock recover signal that the mobile source transmits at the beginning of a packet. A data source transmits the clock signal at the bit rate, and a priority voice source at half the bit rate. The lower rate is used for the priority source so that multiple data sources that pass through a nonlinear component are not confused with a high priority source.

In conventional MSTDM, priority voice sources use CSMA while data sources and the first packet in a

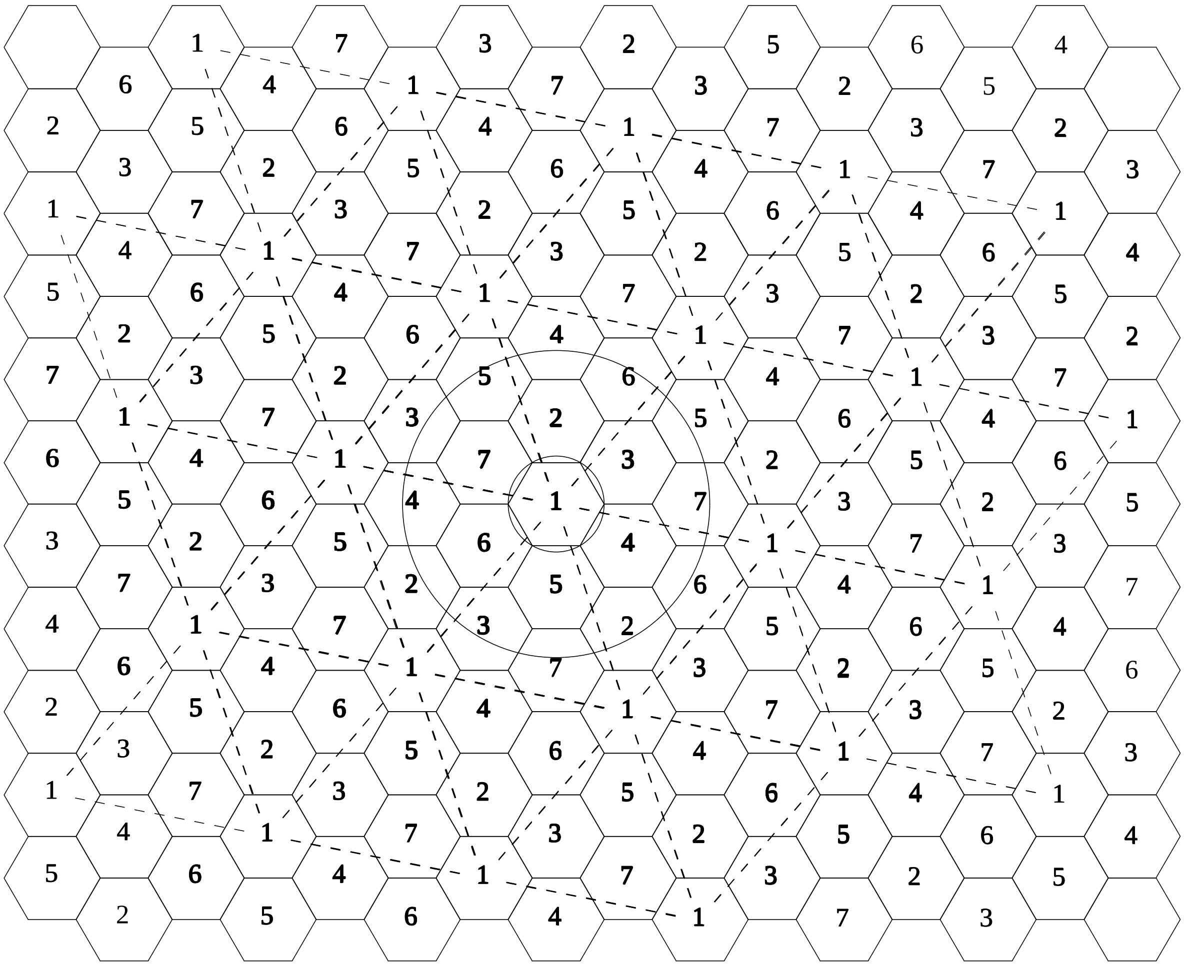

Figure 5. Frequency Plan for CSMA, when mobile units must detect the transmission of mobile units in the same cell

voice stream use CSMA/CD. The service guarantees for voice can be obtained as long as data sources detect collisions with voice sources, but not other data sources. In effect the data sources can implement a CSMA protocol with other data sources, but voice sources must win collisions with data sources. This objective is achieved with the three channel states.

When we use the three state implementation of MSTDM and also use the byte oriented DLE, there is a rare sequence of events that can cause two priority voice sources to collide.[5] In this CSMA-like system the basestation acknowledges correctly received packets so that a data source knows when it has successfully acquired the channel. The acknowledgement is sent as part of the DLE sequences that announces that the channel state has changed from busy to idle. The same acknowledgement is also used to inform two priority voice sources that they have collided and must reacquire the channel. Therefore, the modified MSTDM protocol can continue to operate correctly with occasional collisions between priority voice sources.

Sending acknowledgements for priority voice packets has an added advantage. Cellular systems are typically noiser than wired LAN's and temporary obstructions may cause the channel to fade. The acknowledgement can be used to recover lost packets. When a priority voice packet is not acknowledged, the source can retransmit the packet as a data packet. If the source does not successfully transmit the packet in a "reasonable" amount of time, it assumes that the packet will not arrive at the destination in time to be used in the reconstructed waveform, and discards the packet. A "reasonable" time is a fraction of the period to the source's next scheduled packet.

When data sources and the first packet in a voice sequence collide, the contention resolution algorithm determines whether or not the voice source will acquire the channel quickly enough for the voice source to suppress silent intervals. It has been noted that when these sources are treated equally, and use the contention resolution mechanism used in Ethernet, voice sources cannot acquire the channel quickly enough to avoid clipping a large fraction of the active intervals. In the asynchronous system we use a simple priority mechanism. The first packet in a voice sequence starts to transmit during the interval immediately following a busy channel and the data sources waits until the end of this interval before transmitting. The sources distribute the start of their transmission over the intervals in which they can transmit. The size of the voice access interval is determined by the number of sources that we expect to contend. The number of contending voice sources is derived in reference 5.

4.4 Slotted System

In a cellular network the basestation continuously transmits on the broadcast channel. The mobile units must receive this signal, and it is a natural source of bit timing for their own transmission. By adding a small amount of structure to the broadcast signal it can divide time into slots of any size.

Mobile units transmit at the beginning of a slot. Instead of asynchronously reporting changes in the state of the contention channel, the basestation reports changes at the beginning of each slot. Not every mobile unit receives the start of slot at the same time. The slots must be long enough that a mobile unit can transmit when it detects the beginning of a slot and have that transmission received by the base station before the basestation transmits the beginning of the next slot. A timing diagram that shows the components and duration of a slot is in reference 5. The maximum throughput for slotted CSMA is higher than for asynchronous CSMA[21].

The packets in our system are an integer number of slots. In addition, the time between successive packets from a priority voice source is also a fixed number of slots. Data packets are less than or equal in length to priority voice packets and the overflow area in a voice packet is long enough to transmit any samples that arrive during a voice packet transmission time. With these constraints, we can once again show that priority voice packets never collide,[5] and that the original quality of service is restored to MSTDM.

Once the channel is slotted, other improvements are possible:

— the preempt header in priority voice packets can be eliminated,

— collision detection can be implemented for data contention without significantly increasing the amount of state information, and,

— more efficient contention resolution mechanisms, based on tree searches can be implemented.

The next packet in a priority voice sequence occurs a fixed number of packets after the last. The basestation can prevent a data source from contending for the voice source's minislot by reporting the state of the channel as "reserved". When a priority voice source enters a silent interval the first slot in the packet that it would have taken is wasted because it was reserved. However, two slots at the beginning of each packet during the active interval are saved because the priority source does not send the preempt header.

Collision detection is implemented by having each station send a redundancy check on the first slot of its packet. The basestation changes the channel state to acknowledge that the first slot is successful. When data and voice sources use different redundancy checks, the basestation also determines that the source that has succeeded is a data source or the first packet in an active interval for voice.

Slotted systems can resolve conflicts by partitioning the set of contending sources and using a tree search algorithm, such as those proposed by Hayes[22] or Capetanakis[23]. The access delay for voice sources that enter an active interval is decreased by partitioning the first group of contending sources into initial voice sources and data sources and letting the voice sources contend first[5]. The variance of the access delay of these sources can also be decreased by continuing to partition the tree so that the sources that arrived first, access the system first[24].

When the size of a minislot equals a full voice packet, the operation of this system becomes identical to earlier slotted voice systems with reservations, such as PRMA.[12] In a slotted reservation system the continuation slot is reserved and the voice source does not move. In a movable slot system the minislots are $(1/k)^{th}$ of a voice packet. If the same contention resolution protocol is used in the slotted and minislotted system, then the same number of minislots are used to resolve contention as slots. Contention is resolved more than k times faster, because fewer new sources arrive during the contention interval. In the slotted system a collision uses the channel for a full slot time, while in the minislotted system the channel is occupied for only $(2/k)^{th}$ of a slot. Finally, a data source can transmit for a fraction of a slot. On the negative side, the minislotted system will send control messages on the outbound channel k times as often.

5. Overcrowded Cells

In section 4.2 we noted that the circles that define transmission areas may be placed further apart than the constraints dictate because there is a regular array of hexagons that limits the position of transmission areas. Instead of leaving large portions of the surface unused by frequency bands, the radius of the transmission circle can be increased. As the radius is increased, the service areas overlap, and a larger fraction of the mobile units can be serviced by more than one basestation. In section 5.1 we show how large this fraction becomes.

When a mobile unit can be serviced by more than one basestation, it is possible to redistribute the load out of overcrowded cells and to build hysteresis into the cell boundary to keep mobile units from bouncing back and forth between cells. Overlapping regions is the basis of "soft handoffs" in current CDMA networks.[25] In packet networks we can improve on the "soft handoff" techniques. One way to improve on soft handoffs is to delay handoffs until there is a silent interval. When a handoff occurs during an active interval a glitch occurs in the signal and the mobile unit loses its priority status in MSTDM. Silent intervals can also be used in non-packet networks. Most of the techniques that are unique to packet systems are obtained by treating different priority packets in the same voice stream differently.

MSTDM operates without loss as long as all of the periodic sources use the same packet rate. Instead of treating a voice source as a single packet stream we partition it into two or more periodic streams. For instance, one stream may contain the samples that are needed for intelligible communications, the I channel, and the second stream may contain the samples that provide a higher quality connection, the H channel. There is a cost associated with breaking a voice stream into several different priority streams. Either the packets in each stream are smaller, resulting in a smaller ratio of payload to packet size, or the packet assembly delay is longer. This cost is quantified in section 6.

During normal operation in this example, a mobile unit acquires two periodic channels and sends both packet streams. However, during an overload the mobile unit is instructed by the basestation to treat the I and H channels differently, depending upon the level of oversubscription. The basestation reduces the quality of specific mobile units, rather than reducing the quality uniformly, which makes it possible to define different grades of service, at different prices. We consider three levels of oversubscription, short term, long term, and drastic.

Short term oversubscription lasts on the order of a talk spurt and is caused by too many voice sources becoming active simultaneously. The basestation instructs certain mobile units, or classes of mobile units to send their *H* channel as data instead of priority voice. A mobile unit may send these packets to any of the basestations that can currently receive its packets. Since the packets are all addressed to the mobile unit's agent, it does not matter where they enter the network. The agent merges the packet streams and decides whether to reconstruct the signal from the *I* channel or from all of the packets that it receives before a deadline.

Long term oversubscription lasts on the order of a connection and is caused by more mobile units migrating into a cell than the bandwidth can support. The objective is to redistribute the load by instructing specific mobile units, or classes of mobile units, to redirect their *H* channel or their entire connection to adjacent basestations. The *I* channel should remain with the closest basestation whenever possible. When a station sends the beginning of a talk spurt to the nearest basestation, the basestation is less likely to move out of range before the silent interval.

During long term oversubscription, a protocol is needed to redistribute mobile units. The protocol can use different types of information. A simple protocol could allow mobile units in a congested cell to move to overlapping, less heavily utilized cells. As the utilization in adjacent cells increases, other mobile units in those cells may move further from the congested cell, making it possible for more units to leave the congested cell. This protocol redistributes the load, but may not operate in an optimal manner. In a more sophisticated protocol, basestations could communicate and take into account the number of units that adjacent cells can redistribute and any other congested regions that may be nearby. The basestations could then instruct the proper mobile units to move, even if their cell is currently overcrowded.

Drastic oversubscription is long term oversubscriptions that occurs over a region, rather than a single cell, so that the load cannot be redistributed. The basestation has no option except to shed load. The basestation can instruct specific mobile units to stop transmitting the *H* channel. Only when this measure is insufficient will it become necessary to break connections in overcrowded regions. In section 6 we examine how long we can delay measures that break connections.

5.1 Overlapping Regions

In figure 3, the frequency plan for CSMA/CD, the radius of the circle surrounding a hexagon can be increased from the radius of the hexagon to 1.5 times the radius of the hexagon before the circles corresponding to two cells with the same assignment touch. When the circle circumscribes the hexagon $\alpha = 1$ and when the cells touch, the guard band is zero. While the guard band should not be reduced to zero, it can most likely be reduced from $\alpha = 1$.

Figure 6 shows the effect of allowing the size of the circle to increase. The area at the center of the hexagon that is blank is serviced by one of the three frequency sets, the areas that are striped are serviced by two of the three sets, and the cross hatched areas can use all three frequency sets. In effect, if a mobile unit in a shaded area cannot obtain service from the closest basestation, either because it has failed or because it has too many other users, then the mobile unit can obtain service from one or two other base stations.

The cross hatched area within the hexagon is:

$$A_3 = 6 \left[a * r_c^2 - r_h r_c \sin(a) \right]$$

where, r_h is the radius of the hexagon, r_c is the radius of the circle and $a = \cos^{-1}\left[\frac{\sqrt{3}\, r_h}{2 r_c}\right] - \frac{\pi}{6}$. The striped area is:

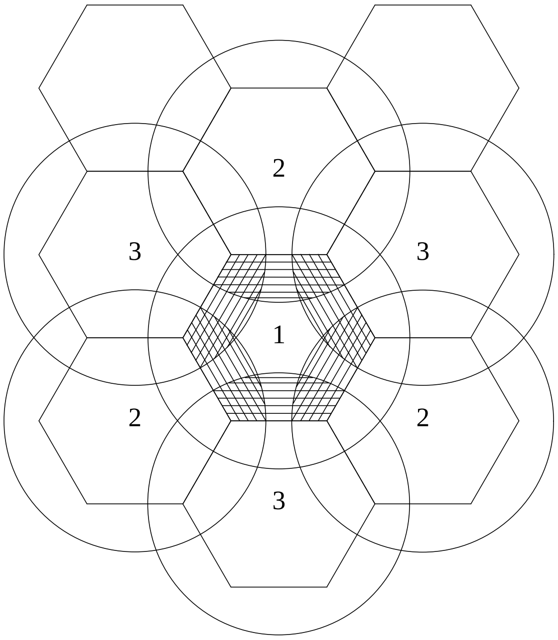

Figure 6. Frequency Overlap

$$A_2 = 6\left[b * r_c^2 - \frac{\sqrt{3}\, r_h r_c \sin(b)}{2} - \frac{A_3}{2}\right]$$

where $b = \cos^{-1}\left[\frac{\sqrt{3}\, r_h}{2 r_c}\right]$.

Figure 7 plots the fraction of the surface that is serviced by one, two or three basestations. As the circles are expanded until two circles that use the same frequency band touch, only 3% of the surface is serviced by a single basestation.

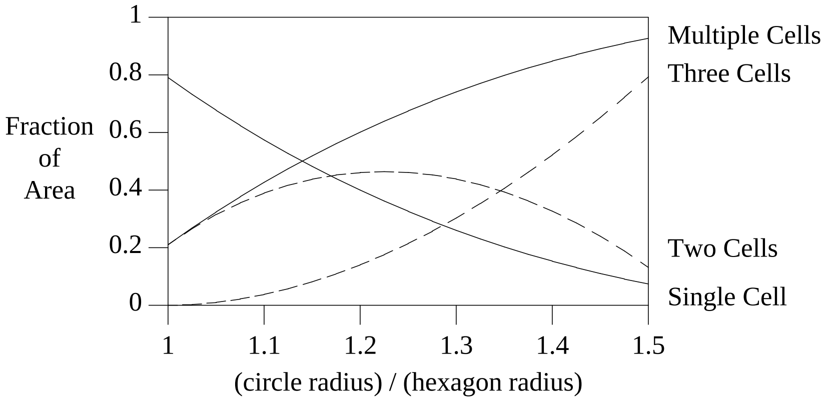

Figure 7. The Fraction of the area in a hexagon that is serviced by one, two, three and two or three basestations, as a function of the radius of the transmission circle from the basestation

6. Performance

There are a large number of parameters that affect the number of voice channels and the the data rate that can be accommodated in a cell. The parameters include the the size of the cell, the bit rate on the channel, the voice coding rate and packet assembly time, the number of partitions of a voice stream, the amount of time required to synchronize reception from a new source, and the size of the header that is used in each packet. In this discussion we set some of these parameters to arbitrary, but representative, values and vary others to determine their effect.

A concern in a packet network that carrys interactive voice is the efficiency of the transmission, defined as part of the voice packet that carrys samples, X_{samp}, divided by the total packet, X_{tot}. There is a trade-off between the efficiency and the delay that is inserted in the voice path. The efficiency is a particular concern in MSTDM because a data packet is constrained to be smaller than a voice packet. Therefore, low efficiency voice transmission also implies low efficiency data transmission.

A voice packet in asynchronous, MSTDM contains:

— a synchronization period, X_{sync}, that is long enough for the receiver to detect the presence of a signal and extract clock,

— a preempt period, X_{pre}, that is long enough for a data source to be informed that a voice source has started transmitting and to stop transmitting before interfering with the information in the voice packet,

— a header, X_{head}, to identify the source and destination,

— the payload, X_{samp}, and

— an overflow area, X_{ov}, that carrys any voice samples that arrive while a voice packet is delayed.

The preempt interval is twice the maximum propagation between a basestation and mobile unit, $\rho_{\max}$, plus the time to wait for a byte boundary, and to transmit the data link escape character, followed by the control character that changes the channel state. $X_{pre} = 2 * \rho \max / r_c + 24$ bits, where r_c is the channel transmission rate. In reference 5 there is a timing diagram that explains the operation of the preempt interval and shows how it may be partially overlapped with X_{sync}. The payload is determined by the rate of the voice coder, r_v, the packet assembly delay, T_V, and the number of partitions of the voice stream, n_p. ($X_{samp} = T_V r_V / n_p$ bits) The overflow area is the solution to $X_{ov} = (X_{sync} + X_{head} + X_{samp} + X_{ov}) r_v / (r_c n_{p})$. The areas in the packet are rounded up to the next byte.

In figure 8 we plot the efficiency of voice packets as we vary six degrees of freedom. The packet assembly delay ranges from 20 to 50 milliseconds, the maximum distance from the basestation to a mobile unit is between a quarter and one kilometer, the channel rate varies from 1 to 100 Mbps, the voice coder rate from 8 to 64 Kbps, and the number of partitions from one to three. For the purpose of this figure, both the synchronization interval and header were set to 64 bits. Data packets are no larger than voice packets. The preempt header and the overflow area are the only overhead bits that are not required in a data packet. Therefore, the efficiency of data packets is only slightly higher than voice packets. The regions with very low efficiencies should be avoided.

In the region that we have selected to study, the cell size is not critical. Therefore, in the ensuing calculations the cell radius is set to 1/2 kilometer. Although 50 msec. packet assembly delay clearly leads to more efficient transmission, 20 msec. has regions with moderate efficiency, is preferable for interactive quality, and is selected for the remaining discussion. The figures indicate that reducing the voice coder rate from 64 Kbps to 32 Kbps or 16 Kbps is probably worthwhile, but further reducing it to 8 Kbps has diminishing returns. Therefore, 16 Kbps and 32 Kbps is selected. There are indications that we may be better off partitioning the bandwidth in a cell into several moderate rate channels, rather than a single high rate channel, but we will retain the full range of channel bit rates.

The next step is to determine the number of voice connections and the data bandwidth that is available at different levels of voice quality. The quality measure is the fraction of the active voice intervals that are clipped because an MSTDM slot cannot be acquired within a packet assembly time. The quality measures that we consider are:

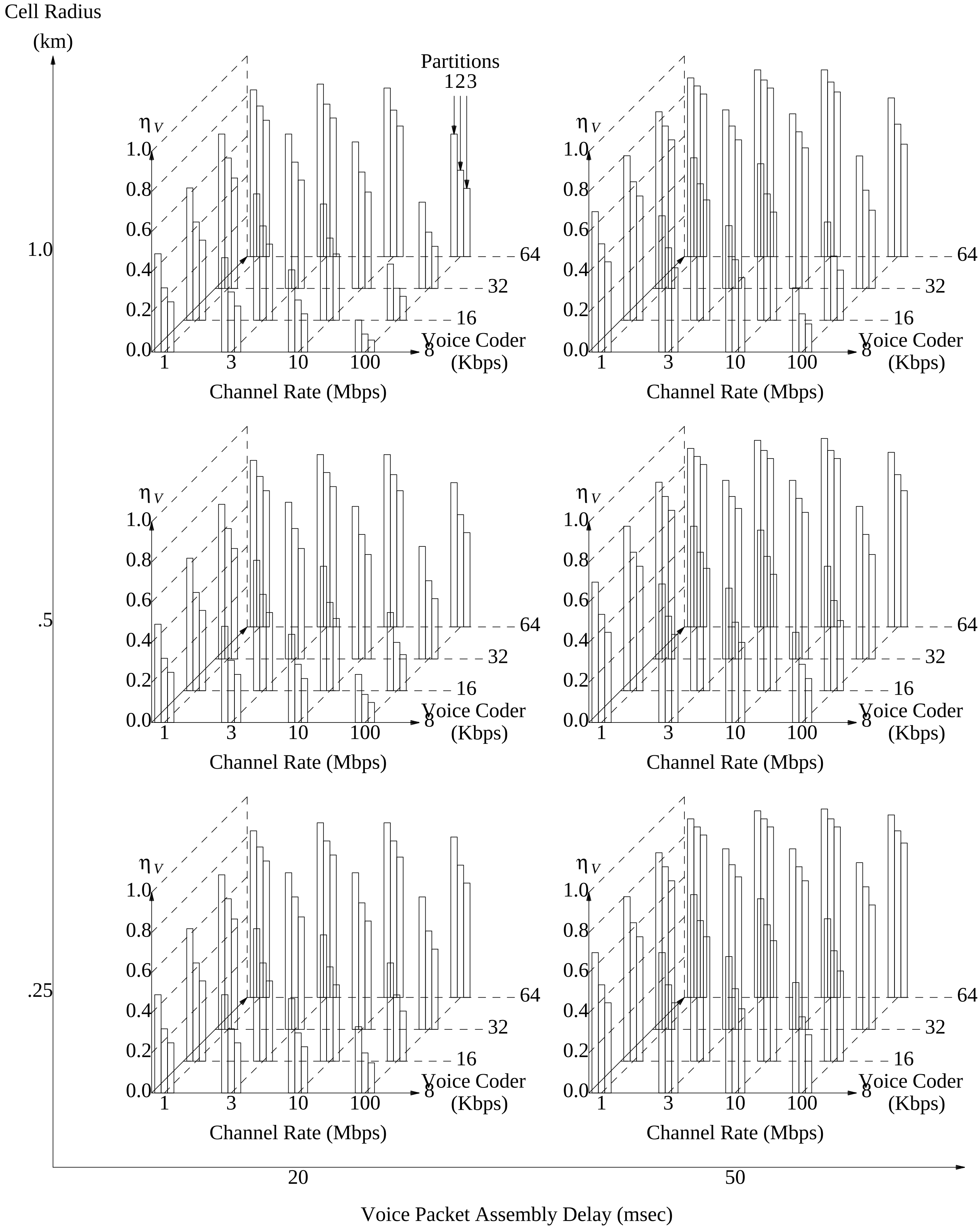

Figure 8. Voice Packet efficiency, η_V = Payload/(Packet Size) as a function of the packet assembly time, radius of a cell, channel transmission rate, and voice coding rate, when the samples from each user are partitioned into 1, 2, or 3, separate packet streams

— 95% of active intervals are transmitted without clipping, which is the requirement that the telephone system used for TASI,

— 99% of the active interval are not clipped, which is a significant improvement on TASI, and

— 100% of the active intervals are not clipped, which is also the number of connections that can be supported in slotted reservation systems that cannot reacquire the channel quickly enough to suppress silent intervals.

We can consider the higher voice quality because the number of connections are larger than in the early systems, which yields a smaller variance in the load, and the unused bandwidth is used for data transmission, rather than being wasted. In order to compare systems with different transmission bandwidths, we normalize the the performance measures. The two measures that we use are the number of voice connections per megabit of transmission bandwidth and the minimum available data bandwidth per megabit of transmission bandwidth.

According to the measurements used to design the TASI system[26,27], the average voice signal is active about 40% of the time, and the different conversations that share the inbound wireless channel are independent. Therefore, the number of active connections, n_a, is a binomial distribution with mean $\mu = .4n$ and variance $\sigma^2 = .24n$, where n is the number of connections in the cell. The average silent interval is 1.7 seconds and the number of new active voice intervals per voice packet period, n_f, is a Poisson distribution with $\lambda = (1.7/T_V)(n-n_a)$. In order to avoid clipping an active interval, the first packet in an active interval must acquire the channel within T_V. In general, the likelihood of the first packet acquiring the channel depends on the fraction of the bandwidth that is not used by currently active voice connections and the contention from data sources. The priority mechanisms that we have described for both asynchronous and synchronous systems permit us to ignore contention from data sources.

Calculating the probability of clipping is a complex process that depends upon the distribution of busy and idle intervals on the channel and the specific implementation of the contention resolution algorithm. The distribution of idle and active intervals in MSTDM is dependent upon the entire past history of delays. We can obtain an approximate measure of voice connections by realizing that there are very few voice arrivals per T_V, particularly when most of the connections are active and the likelihood of not resolving contention is greatest. Contention is most difficult to resolve if there is a busy period during which all of the new voice sources arrive followed by a resolution period in which they are all trying to transmit. With the parameters that we are considering, when the system is loaded to the point where clipping may occur, n_f is seldom more than 5 or 6. With this small number of contending sources, contention is almost always resolved in $2*n_f$ voice packet transmission times.

The maximum number of voice packets that are transmitted during T_V is $n_{\max} = r_c T_V / X_{tot}$. We assume that when $n_a + 2*n_f \le n_{\max}$ none of the new sources are clipped, and when it is greater, $\lfloor n_f - (n_{\max} - n_a)/2 \rfloor$ are clipped. (Where $\lfloor x.y \rfloor$ increases $x.y$ to $x+1$ when $y \ne 0$ and $\lceil x.y \rceil$ reduces $x.y$ to x.) We calculate a starting of value of n, at a given quality level, by letting $n_a = n_{\max} - 1$ and approximating the binomial distribution for n_a by a normal distribution. The starting value of n is th number of sources that has fewer than n_a active sources 95% of the time. For instance, at the 95% level, n is the solution of the quadratic equation $n_{\max} = .4n + 1.645\sqrt{.24n}$. This is an upper bound on n, since it allows fro a single arrival 95% of the time, but does not take into account multiple arrivals. We decrement n by one at a time until:

$$\sum_{n_a=0}^{n_{\max}} \sum_{n_f=\lceil (n_{\max}-n_a)/2 \rceil}^{n} \frac{\lfloor n_f - (n_{\max} - n_a)/2 \rfloor}{n_f} Pr(n_a/n) Pr(n_f/(n_a,n)) < 1 - P_Q$$

where $Pr(n_a/n)$ and $Pr(n_f/(n_a,n))$ are the binomial and Poisson distributions defined earlier, and P_Q is the fraction of the active intervals that are not clipped.

The iterative procedure is used to determine the number of voice connections that can be supported for P_Q equal to .95 and .99 in figure 9. The number of voice connections for $P_Q = 1$ is $n_{\max}$, which is the level at which every voice connection can be active. The minimum data rate is the bits per second per megabit of transmission that is still available to data sources when the maximum number of voice sources are in the cell. The minimum data rate at 100% assumes that silent intervals are suppressed. When silent intervals are not suppressed, the minimum data rate is 0, rather than 60%.

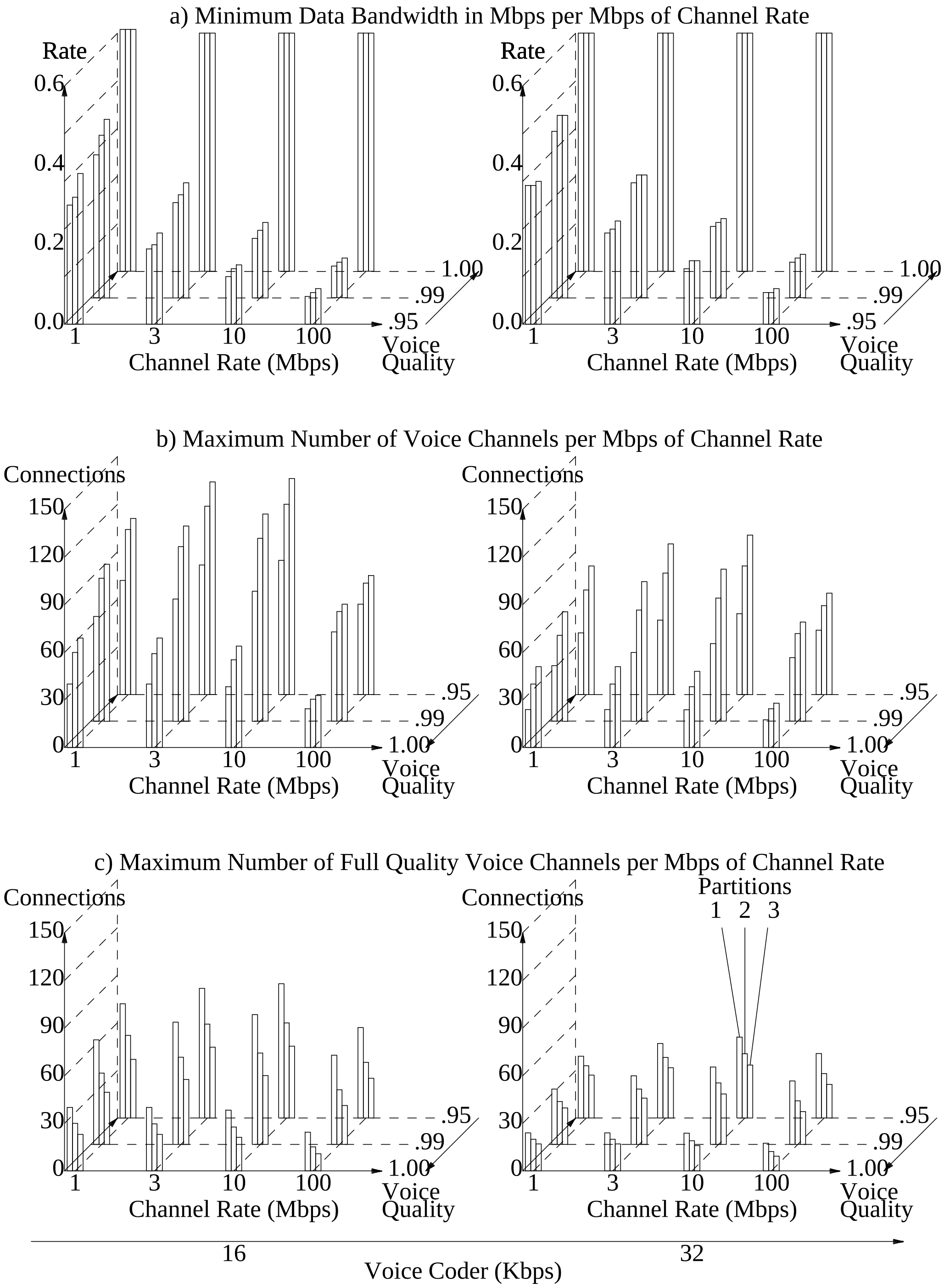

Figure 9. System Throughput: a) The channel rate that is available for data with the maximum number of voice connections; b) The maximum number of voice channels when only the most important partition is transmitted; c) The maximum number of voice channels when all of the partitions are transmitted.

Figure 9.c shows the maximum number of full quality voice connections that can be supported at a particular clipping level. This is the level at which the network must start shedding or redistributing load.

Note that the level decreases as a voice stream is partitioned into more pieces, because the ratio of overhead to payload increases. Therefore, in a partitioned system connection quality is reduced before an unpartitioned system must shed connections. In figure 9.b we show the levels at which all of the systems must shed load. As expected, the unpartitioned system is at the same level as in 9.c and the partitioned systems function at higher levels. If the system is full to capacity with voice only 1% or 5% of the time, then most of the time there is bandwidth that is still available for data. The residual bandwidth can be appreciable, as plotted in figure 9.a.

For the set of parameters that we have selected, the number of voice connections per megabit seems to peak at network rates between 3 Mbps and 10 Mbps. At these two channel rates the number of voice connections per megabit is about the same, but the residual bandwidth available for data is much higher in the 3 Mbps systems. Based upon these figures, there are regions where we may be better off dividing the bandwidth in a cell into several channels that are shared separately. For instance, if we can transmit 30 Mbps in a single channel or 3 Mbps in 10 channels, we should use 10 channels and make it the basestations responsibility to shift load between the channels.

7. Conclusion

In order to obtain the full advantage of applying packet techniques to a mobile network, the network must contain a wired and wireless segment. The wired segment makes it possible keep the connection between the mobile unit and the current communications network (WAN) stationary. The path changes that occur when mobile units change cells is performed in the wired portion of the mobile network (MAN).

When mobile units change cells, stateless packet switching reduces the amount of work required relative to stateful circuit switching. A mobile units interface (agent) between the MAN and WAN is in a fixed location. The packets from the mobile unit are addressed to the agent and conventional packet switching techniques get the packets to the agent independent of where the packets enter the wired network. There is no work involved in handoffs between cells in this direction. In the other direction, the agent encapsulates packets to the mobile unit inside packets that are addressed to a basestation. The basestation changes when the mobile unit moves between cells and this change is reported to the agent. The work involved in changing the address for the encapsulation packet is less than the work that would be needed to tear down a circuit to a basestation and set up a circuit to a new basestation.

On the wireless portion of the mobile network packet techniques provide a means of conserving bandwidth and dealing with overcrowded cells. Packets are only transmitted during active speech intervals in a voice connection, which results in a 60% reduction in bandwidth. Speech is partitioned into separate components that are required for intelligibility and components that add to the quality. When a cell is overcrowded the quality of selected mobile units is reduced, before denying service to any mobile units. Applying packet techniques on the outbound channel is a simple multiplexing function. The quality of service on the inbound channel is maintained using a variant of MSTDM. By taking advantage of the characteristics of the network, the variant requires less bandwidth than previous applications of this protocol.

A separate packet technique is used for a common control channel. Mobile units receive this channel independent of their location. This channel is used to place calls to inactive mobile units.

REFERENCES

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

[2] N. Abramson, "The Throughput of Packet Broadcasting Channels," IEEE Trans. on Commun., vol. COM-25, no. 1, Jan. 1977, pp 117-128.

[3] J. Edney, "802.11: The New Wireless Standard," Data Communications, pp 77-86, 1995.

[4] 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.

[5] N. F. Maxemchuk, "Applying MSTDM to Packet Voice and Data in Cellular Radio Systems," ICNP '98, Austin Tx, Oct 14-16, 1998, pp. 309-318.

[6] F. E. Ross, "FDDI-a Tutorial," IEEE Comm. Mag., May 1986, vol.24, no.5, pp. 10-17

[7] R. M. Newman, Z. L. Budrikis, J. L. Hullett, "The QPSX Man," IEEE Comm. Mag., Vol. 26, No. 4, April 1988, pp. 20-28.

[8] N. F. Maxemchuk, "Regular Mesh Topologies in Local and Metropolitan Area Networks," AT&T Technical Journal, Sept. 85, vol. 64, no. 7, pp. 1659-1686.

[9] 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.

[10] N. F. Maxemchuk, R. Krishnan, "A Comparison of Linear and Mesh Topologies - DQDB and the Manhattan Street Network," IEEE JSAC, Oct. 1993, vol. 11, no. 8, pp. 1278-89.

[11] J. T. Brassil, A. K. Choudhury, N. F. Maxemchuk, "The Manhattan Street Network: A High Performance, Highly Reliable Metropolitan Area Network," Computer Networks and ISDN Systems, Mar. 1994.

[12] 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.

[13] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications," Internet Engineering Task Force, Audio-Video Transport Working Group, Jan.1996.

[14] C. E. Perkins, "Mobile IP," IEEE Commun Mag, May 1997, vol. 35, no. 5, pp. 84-99

[15] D. J. Goodman, A. A. M. Saleh, "Local Aloha Radio Communication with Capture and Packet Buffers," Proc. Globcom '85, Dec. 2-5 1985, New Orleans, Louisiana, pp. 984-990.

[16] 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.

[17] 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.

[18] P. T. Brady, "A Technique for Investigating On-Off Patterns of Speech," BSTJ, Vol. XLIV, No.1, Jan. 1965, pp 1-22.

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

[20] A. R. Cohen, J. Heller, A. J. Viterbi, "A New Coding Technique for Asynchronous Multiple Access Communication," IEEE Trans. on Commun Tech., vol. COM-19, no. 5, Oct 1971, pp. 849-855.

[21] 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.

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

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

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

[25] A. J. Viterbi. A. M. Viterbi, K. S. Gilhousen, E. Zehavi, "Soft handoff extends CDMA cell coverage and increases reverse link capacity," IEEE JSAC, Oct. 1994, 1281-1288.

[26] P. T. Brady, "A Model for Generating On-Off Speech Patterns in Two-Way Conversation," BSTJ, Sept. 1969, pp. 2445-2472.

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