A New Adaptive MAC Layer Protocol for Broadband Packet Wireless Networks in Harsh Fading and Interference Environments

Anthony S. Acampora
Center for Wireless Communications
University of California, San Diego
9500, Gilman Drive,
La Jolla, CA 92093
acampora@ece.ucsd.edu

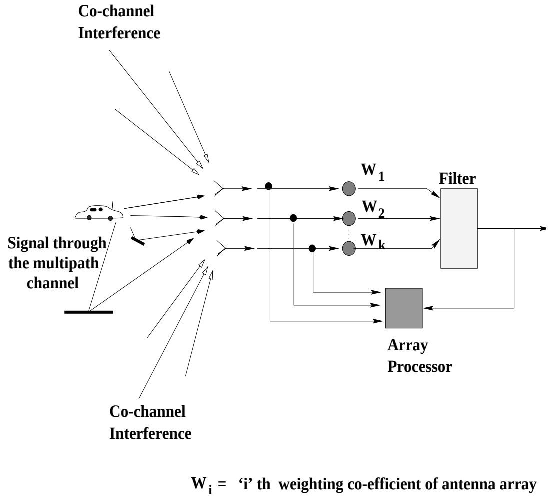
Srikanth V. Krishnamurthy
Information Systems Laboratory
HRL Laboratories,
3011, Malibu Canyon Road,
Malibu, CA 90265
krish@wins.hrl.com

Abstract

A new medium access protocol is proposed for sharing a high speed radio channel among a number of small wireless *packet access units*, some of which may be stationary and some of which may be within moving vehicles. Such a system could provide fixed point, pedestrian and remote users with wireless access to CPU and database resources of an underlying ATM wireline network, essentially extending the ATM bandwidth-upon-demand interface directly to the wireless units and enabling delivery of multimedia services (albeit at the lower peak rate afforded by the radio channel). A primary goal of the proposed medium access protocol is the pre-delivery of a signal from each packet access unit as needed to rapidly compute the weights needed by a base station's adaptive array processor or a space-time processor, thereby protecting the packet flow in each direction from the effects of both multipath propagation and adjacent channel interference arising in neighboring radio cells. An impairment-robust direct sequence spread spectrum-based polling signal is invoked to stimulate a pilot tone from a given remote immediately prior to packet transfer in either direction, thereby permitting the base station to determine a good set of antenna element combining or power splitting weights to be used for that packet. Reasonable approximations are invoked to study the performance of the proposed protocol, and link utilization efficiency and average message delay are found. By proper choice of protocol parameters, a radio resource utilization efficiency of about 95 % is readily achieved. Accuracy of the approximations is confirmed by extensive computer simulations.

1 Introduction

Frequency selective fading and co-channel interference are two well known impairments which distinguish cellular from wireline communications. The first is caused by multipath propagation and the second, caused by frequency reuse among the various radio cells, demands some minimum geographical separation between any two cells sharing a common frequency. The use of spatially diverse antenna array elements at the base station is a well known strategy for abating both multipath fading and co-channel interference [1]- [4]. The adaptive array would help abate the time-varying impairments of the radio channel to produce, with very high probability, the low Bit Error Rate (BER) needed by ATM. In the remote to base direction, the signals appearing at the various elements are phase and amplitude weighted and summed such that some objective function is optimized [5]. The same weights may be used in the *reciprocal* base to remote direction. Figure 1 depicts a base station with an array antenna. The array processor computes the weights W_j and tunes the receiver so as to abate the fading and interference effects of the channel.



Co-channel Interference
W1
W2
Wk
Filter
Signal through the multipath channel
Array Processor
Co-channel Interference
W_i = 'i' th weighting co-efficient of antenna array

Figure 1: A base station equipped with an array antenna

In this paper, we present a wireless medium access protocol for enabling packet-based communications between each remote terminal (or packet access unit) within a radio cell and that cell's base station, designed to be used in conjunction with a base station antenna array. Time-Division Duplexing (TDD) is used, with a single high speed channel shared by the base station and all remote units within a cell. Such a scheme might be useful for extending broadband wireline services to wireless stationary, pedestrian, and remote users, supporting bandwidth upon demand by means of ATM transport [6].

If we imagine an orderly time sequence of time-duplexed packets flowing between the remote terminals and the base station, then the medium access protocol must support the base station's ability to determine, on a packet by packet basis, the correct set of receive and transmit weighting coefficients to be used. Since the multipath and interference environments are time varying, the weighting coefficients for each packet access unit must be continually updated. It is to be noted that in this paper, our goal is to provide a medium access control protocol which allows the base station to obtain some kind of training sequence or pilot tone. The training sequence or pilot tone enables the antenna array to compute a set of weighting co-efficients or the combined set of antenna weighting coefficients and transversal equalizer coefficients if the bandwidth is wide and space-time processing is required. The paper however, does not discuss what the training sequence ought to be or what hardware processing capabilities are needed. Thus, whether the processing required is merely spatial, or whether joint spatial and temporal processing is required, the medium access control protocol enables the array antenna to obtain the training sequence to compute its weighting co-efficients with as low an overhead as feasible.

In a previous publication [7], a modified token protocol was described which allowed the base station, using a known set of weights for each remote, to sequentially prompt a response from each remote. This response would take the form of either an information packet (if the remote had queued information awaiting transmission) or an unmodulated pilot tone (otherwise). Either response allows the base station to observe that remote's signal as it appears at each antenna element, and to update the weighting coefficients accordingly. The complex conjugate coefficients would then be used in the base-to-remote direction and the polling frequency would be sufficiently high such that the base station is always in possession of the correct set of weights for each remote. With this scheme, overall channel utilization efficiency degrades as the polling rate increases since each remote must be polled once per polling cycle and time spent polling is unavailable for delivery of user information. It was shown in [7] that, for typical pedestrian-oriented system parameters, a polling cycle of 50 milliseconds yields an overall channel utilization efficiency in excess of 80 %.

While a 50 millisecond polling cycle may be adequate for pedestrian traffic, it is far too long for vehicular traffic as the following example illustrates. Consider a vehicle traveling at a velocity of 60 mph. In 50 msec, the vehicle has traveled 4.4 ft. At a carrier frequency of 900 MHz, the wavelength is approximately 1 ft, and we see that a distance corresponding to 4.4 *wavelengths* is traveled per polling cycle. Since multipath fading is caused by the destructive interference of many reflected rays, the phase angles associated with each ray will be totally randomized when the base station next sends the token, 4.4 "wavelengths" later, and the currently stored set of weights will be woefully outdated. Needed is a medium access protocol that supports much faster array adaptation.

The high speed radio channel is assumed to be shared by a plurality of small remote terminals which may be stationary (fixed point services), carried by pedestrians, or used in moving vehicles, and which can access the CPU and database resources of the wireline network. In the basic mode of interaction, the remote would make a short inquiry to the base station. The response from the wireline resources might include a long data file, an image etc.

In the medium access protocol proposed in this paper, the base station again periodically polls each remote; the polling signal uses multipath and interference-robust direct sequence spread spectrum modulation [8] [1]. In response to a poll, the remote either replies with information at some pre-determined future time (if it has one or more packets to send) or is silent. If a packet is sent, a preamble is included to be used by the base station for rapid acquisition of the correct array weights for that packet. The base station may also command a given remote to transmit a pilot tone at some predetermined future time; in response to this pilot, the base station rapidly computes a set of weights to be used immediately for sending information to that remote. With such a scheme, the delay between acquisition and use of the weighting coefficients is extremely short. By way of example, if a message contains, say, 10 ATM cells and the channel data rate is 20 Mbps, then the message length is less than 250 microseconds. Within this time interval, a vehicle traveling at 60 mph has moved only about 0.25 inches. Thus, at a carrier frequency of 900 MHz, the phase variation of each multipath ray over the duration of the message is negligible.

We develop an analytical model from which approximations for the effective utilization efficiency and expected message delay may be found. Results for typical system parameters indicate that a utilization efficiency as high as 95 % may be achieved. It is also observed that the expected message delay is relatively independent of the number of remotes in the cell.

2 The Proposed Medium Access Protocol

Consider a microcell containing a single base station and a plurality of remote packet access units. The communication between the base station and remotes is assumed to be time-division duplexed over a single high speed radio channel. As described in Section 1, the base station is equipped with an adaptive array antenna to abate multipath fading and co-channel interference and since these impairments may rapidly change, the medium access protocol is designed to provide the base station with a pilot signal sent by a remote, prior to information transfer to or from that remote.

A timing or framing diagram of the proposed medium access protocol appears in Figure 2. The frame consists of two segments: a polling segment and a data segment. The polling segment contains N polling slots, one for each remote within the radio cell. The data segment contains $J > N$ information transfer slots and is further subdivided into two subfields. Subfield 1, containing N slots, is used for transferring *information requests* from remote-to-base, and may, under certain circumstances, be used to transfer *information replies* from base-to-remote. Subfield 2 is used exclusively for base-to-remote information replies.

The polling segment is used to signal an intent to send an information request (remote-to-base) or reply (base-to-remote). Multipath and interference-robust direct sequence spread spectrum signaling is used during the polling slots, unprotected by the array. Each polling slot is *owned* by a particular remote and consists of two parts (Figure 2). In the first part of its polling slot, if the *owner* remote has a queued information request to send, it sends in its *own* spreading sequence which occupies the entire channel bandwidth. (Note that, during the data segment, information is sent at a rate which consumes the entire spectrum; the chip rate during the polling segment is therefore equal to the data rate during the data segment, and a spreading sequence of say, 200-300 chips should be adequate to distinguish the remotes, suppress the interference, and abate multipath. Note further that we do not consider the practical issues of synchronizing the correlator needed at the base station and each remote, or implementation of the requisite matched filter if synchronization is to be avoided.) The base station immediately recognizes this polling signal, and acknowledges receipt using the same spread spectrum signal in the second half of that polling slot. Corresponding to each polling slot, there exists a data slot in Subfield 1 of the data segment. The acknowledgement from the base station notifies the remote that the corresponding data slot in Subfield 1 is available for that remote to send its information request. Just prior to that particular data slot, the remote would then send a short pilot tone (Figure 2) to be used by the base station for rapid acquisition of the currently correct set of antenna array combining weights to be used for that remote. If the remote does not have a queued information request, it does not send its spread spectrum sequence in its polling slot, thereby notifying the base station that the corresponding data slot in Subfield 1 of the data segment is free and may be used for transfer of an information reply to any remote. To do so, the base-station transmits the spread spectrum sequence of the chosen remote, using the base-to-remote portion of the polling slot; this notifies the remote that the corresponding data slot in Subfield 1 of the data segment has now been allocated to it for receiving a base-to-remote information reply. The chosen remote, then, sends a pilot tone just prior to that slot, allowing the base station to acquire the correct set of

[1] We assume that the base station has prior knowledge of the spread spectrum codes and can use a matched filter receiver for capturing the intended users transmission. In contrast, circuit switched CDMA systems might require code acquisition. For a discussion on spread spectrum communications and how the underlying communications including code acquisition is done, one might refer to [9].

antenna array weights to be used for that information reply.

As described, each slot of Subfield 1 serves two purposes. If a slot *owner*, say the *j* th, has a queued information request, the *j* th slot in Subfield 1 is used to transfer that request. Otherwise, the base station may use that slot for transferring an information reply packet to *any* particularly chosen remote. In Subfield 1, data packet transfer must be preceded by a pilot tone since the base station must acquire the correct set of antenna array weights to be used for each packet. By contrast, the slots of Subfield 2 are used exclusively for base-to-remote information replies. Here, it is possible for the base station to schedule a long multi-slot reply to one particular remote, and it will not, therefore, need to re-acquire the antenna weights prior to each slot. Rather, the transfer of information can continue without pilot tones until the long message ends. In such a case, the base station would *asynchronously* send polls to the recipient of the next long message, stimulating a pilot tone for re-acquisition purposes as needed (See Figure 2). The base station would then rapidly acquire the weighting coefficients for its antenna array and transfer of the long information reply would continue. Note that it is possible to embed in each time slot a signaling segment to be used to indicate the beginning, end, or continuation of a message.

It is to be noted that although the model assumes that the information replies are in the base-to-remote direction, it is possible that the replies may actually refer to long file transfers in the remote-to-base direction. The remote would indicate the length of the file in its request and the base station can allocate the requisite capacity in the data segment for file transfers. Thus, the medium access protocol supports transfer of bursty data in both the base-to-remote and the remote-to-base directions.

3 Performance Analysis

In this section, we present an analytical model for the proposed medium access protocol and derive close approximations for the expected message delay and *channel utilization efficiency*. The frame size can be appropriately chosen so that the latter is maximized to achieve *maximum utilization efficiency*. We consider the following traffic model. Each remote-generated information request message consists of a fixed number of ATM cells and fits exactly into one data slot of the frame. A remote can generate a new information request only if it has no outstanding requests, i.e., it can generate, at most, one new request per frame. We define p to be the probability that a given remote generates a request packet in some given frame, and $q = 1 - p$ is the probability that that given remote does not generate a request packet in that frame.

We assume that each base-to-remote information reply message contains a geometrically distributed number of packets, and that the replies are sent to the remotes on a first-come-first-served basis. We also assume that there are N remotes within the cell and that the data segment contains $J > N$ slots (Figure 2).

3.1 Link Utilization Efficiency

The *utilization efficiency* is defined as the fraction of the time that useful information is being transmitted over the channel, where, by definition the useful information consists only of information requests and replies. Thus, the utilization efficiency is simply the ratio of time consumed by data slots carrying useful information to the total duration of the frame.

To calculate the utilization efficiency, we first find the average number of asynchronously inserted polling slots and pilot tones in Subfield 2 of the data segment [2].

We assume that the number of packets in a message is geometrically distributed with parameter β, i.e.,

$$P(\text{Message has } u \text{ packets}) = (1-\beta)\beta^{u-1}; \quad u \geq 1. \tag{1}$$

Let $M(z)$ be the probability generating function for the number of packets in a message. Then,

$$M(z) = \sum_{u=1}^{\infty} \beta^{u-1}(1-\beta)z^u \tag{2}$$

We note that since the number of packets in a reply message is geometrically distributed, the distribution of the number of packets remaining to be transmitted after a poll is the same as the overall distribution of the number of packets in the message.

Differentiating Equation (2) and setting $z = 1$ yields $\overline{M}$. The average number of polls in Subfield 2 is therefore given by:

$$\delta = \left\lceil \frac{J-N}{\overline{M}} \right\rceil \tag{3}$$

where $\lceil x \rceil$ denotes the closest integer greater than or equal to x. It is to be noted that the size of each polling interval in Subfield 2 of the data segment is equal to half the size of a polling slot since a polling signal is needed only in the base-to-remote direction.

Further, if we choose a slot at random, the probability that it would not contain a remote-to-base request is given by $\gamma = \frac{J-N}{J} + \frac{N}{J}q$. Thus if P_0 is the probability that the base station queue is empty, the utilization efficiency is given by,

$$\rho = \frac{Jn_1[(1-\gamma) + \gamma(1-P_0)]}{Jn_1 + (N + \frac{\delta}{2})n_2 + (N+\delta)n_3} \tag{4}$$

where $n_1, n_2,$ and, n_3 the number of ATM cells in a data slot, the number of ATM cells in a polling slot and the number of ATM cells in a pilot tone respectively.

For a given frame length J, the fraction of information bearing slots will be maximized when P_0 is at its minimum. Consequently, for a given frame length, the *highest utilization efficiency* is achieved when P_0 is minimized. Then, by optimizing with respect to J, the minimum value of P_0 can be made zero [3], thereby yielding the *maximum utilization efficiency*, which is given by :

$$\eta = \frac{Jn_1}{Jn_1 + (N + \frac{\delta}{2})n_2 + (N+\delta)n_3} \tag{5}$$

3.2 Mean Delay Analysis

In this subsection we compute an analytical expression for the mean delay which refers to the average time between the moment that a remote's information request reaches the base station and the moment when it completes the reception of the corresponding information reply from the base station.

Let us consider an arbitrary frame, say the "ν" th frame, and let us further consider the instant when the n th slot of its data segment ends. If that slot is in Subfield 2 of the data segment, the number of packets in the base station's queue is decremented by one. On the other hand, if that slot is in Subfield 1 of the data

[2] These slots are actually very few in number and do not change the results by much. However, an approximate estimate is computed.

[3] To be discussed in Section 4.

segment, the number of packets in the queue will be decremented by one if the particular remote (say the ith) *owning* that slot does not send an information request in that frame. If the slot owner sends an information request, then the number of packets in the base station's queue will be incremented by m_i, where m_i is a random variable representing the number of packets in a message requested by the i th remote. For simplicity, we assume that all information reply messages contain an identically distributed number of packets, i.e., m_i has a probability generating function (*pgf*) which is denoted by $M(z)$, which is the same for all remotes.

For the medium access protocol described in Section 2, the number of slots in Subfields 1 and 2 are permanently fixed at N and $J - N$, respectively. We consider an analytically simpler approximation, whereby the number of slots in each subfield is a random variable, with the expected values of each equal to the actual *fixed* number of slots, N and $J - N$, respectively. Then, we model the two subfields as the states of an alternating renewal process [11]. Subfield 1 of the data segment is the first state of the alternating renewal process, and the Subfield 2 is the second state. The probability that the next slot belongs to Subfield 1 is $\frac{N}{J}$ (the average fraction of the time spent in state 1), and the probability that it belongs to Subfield 2 is $\frac{J-N}{J}$ (the average fraction of the time spent in state 2). These approximations are very good for estimating first order delay statistics since the first order expectation values approach the constant fixed values in the steady state. The accuracy of the approximations is further validated by simulation studies of the actual protocol.

Let us consider the boundary between the nth and the $(n+1)$ st slots. Let Y_n represent the number of packets in the base station queue at the beginning of the n th slot. We define:

$$U(Y_n) = \begin{cases} 0 & \text{if } Y_n = 0 \\ 1 & \text{if } Y_n > 0. \end{cases} \tag{6}$$

Then, if the current slot corresponds to state 1 and if there was no request from the corresponding remote,

$$Y_{n+1} \quad = \quad (Y_n - U(Y_n)). \tag{7}$$

Also, if there was a request from the corresponding remote, then

$$Y_{n+1} \quad = \quad (Y_n + m). \tag{8}$$

where m is the number of packets in the message requested by that remote.
If the current slot corresponds to state 2 :

$$Y_{n+1} \quad = \quad (Y_n - U(Y_n)). \tag{9}$$

Applying the methodology used in chapter 4 of [12], the generating function of the random variable Y_{n+1} can be expressed as

$$\begin{aligned} Y_{n+1}(z) &= E(z^{Y_{n+1}}) = E(z^{Y_n - U(Y_n)}).\gamma \\ &+ E(z^{Y_n + m}).(1-\gamma), \end{aligned} \tag{10}$$

where $\gamma = \frac{J-N}{J} + \frac{N}{J}q$. This follows directly from finding $E(z^{Y_{n+1}})$ conditioned on each of the cases represented by Equations (7), (8) and (9), and averaging over the probabilities of these cases. Equation (10) can then be reduced to

$$\begin{aligned} Y_{n+1}(z) &= [P_0 + z^{-1}(Y_n(z) - P_0)].\gamma \\ &+ Y_n(z)M(z)(1-\gamma). \end{aligned} \tag{11}$$

where P_0 is the steady state probability that the base station queue is empty. In the steady state, as $n \to \infty$, $Y_{n+1}(z) = Y_n(z) = Y(z)$, and hence, Equation (11) can be thus expressed as

$$Y(z) = \frac{P_0(1 - z^{-1})\gamma}{1 - M(z)(1-\gamma) - z^{-1}\gamma}. \tag{12}$$

Since $Y(z)$ is a generating function, $Y(1) = 1$, and using this it can be easily shown that

$$P_0 = 1 - \frac{\overline{M}(1-\gamma)}{\gamma}. \tag{13}$$

From $Y(z)$ we can also find $\overline{Y}$ which yields the average number of packets in the queue.

We shall now find an approximate expression for the expected message delay.

To find this, we first have to find the distribution of the time taken to transmit a message, once it gets to the head of the base station's queue. If l is a random variable denoting the total time taken to transmit an information reply message, once it gets to the head of the base station queue, then,

$$E(z^l/\text{message has } u \text{ packets}) = E(z^{h_1 + h_2 + \ldots + h_u}), \tag{14}$$

where h_i is the time (in number of slots) between the transmission of the $(i-1)$ th and the ith packet of the information reply message. We shall call this time the *holding time* for packet i.

The holding time is the time for which a packet, on reaching the head of the base station's queue, remains in the queue before transmission. Using the alternating renewal process model, we find an approximate expression for the probability generating function of h_i. With this model, all packets have independent and identically distributed holding times. We denote the *pgf* of this holding time distribution by $H(z)$, which is given by:

$$\begin{aligned} H(z) &= z\frac{J-N}{J} + (\sum_{u=1}^{N-1} z^u (1-q)^{u-1} q \\ &+ \sum_{u=N}^{\infty} z^N (1-q)^{u-1} q)\frac{N}{J} \\ &= z\frac{J-N}{J} + (\frac{zq(1-(z(1-q))^N)}{(1-(1-q)z)} \\ &+ z^N(1 - \frac{q\{(1-(1-q)^N\}}{q}))\frac{N}{J}. \end{aligned} \tag{15}$$

The first term accounts for the holding time a packet experiences when it is in the first state of the alternating renewal process and the second term for the holding time it experiences when it is in the second state. Since the holding times for the packets are independent, Equation (14) can be written as

$$E(z^l/\text{message has u packets}) = [H(z)]^u, \tag{16}$$

Thus using Equations (1) and (16),

$$\begin{aligned} E(z^l) &= \sum_{u=1}^{\infty} (H(z))^u \beta^{u-1}(1-\beta) \\ &= \frac{H(z)(1-\beta)}{1-\beta H(z)} \equiv L(z). \end{aligned} \tag{17}$$

The message delay contains two components, i.e., the message transmission delay and the queuing delay [4]. Upon arrival, let the message find k packets in the queue (including the one being served). Then, the time for the message to reach the head of the queue is equal to the transmission time of the k packets, and the characteristic function of the queuing delay is given by

$$\begin{aligned} Q(z) &= \sum_{i=0}^{\infty}(H(z))^i P(i \text{ packets in the queue}) \\ &= Y(H(z)). \end{aligned} \tag{18}$$

Also, since the queuing delay is independent of the message transmission delay, the probability generating function or the *pgf* of the message delay, $D(z)$, is given by the following expression:

$$D(z) = Q(z)L(z) = Y(H(z))L(z). \tag{19}$$

Differentiating the above expression and setting $z = 1$ yields the average message delay, given by:

$$\overline{D} = \overline{L} + \overline{YH}. \tag{20}$$

In the above analysis, it is assumed that the probability that a particular remote does not generate a request in a particular frame is some constant q. Although the number of data slots in a given frame is fixed, the amount of overhead due to polling and pilot tones may vary from frame to frame since the base station may asynchronously insert polls in Subfield 2 of the data segment, if necessary, as explained in Section 2. However, the number of these asynchronously inserted polls is negligible since they are inserted only if the base station must initiate transmission of a new message, and we assume that information reply messages, on average, contain a large number of data packets. Hence, we may consider q to be a constant every frame.

4 Results and Discussion

For illustrative purposes, we assume that a data slot is of duration equal to 3 ATM cells, a polling slot is of duration 1 ATM cell (that is, the spreading sequence used in the remote-to-base and base-to-remote direction each contain 212 chips), and a pilot tone is of duration 1/2 ATM cell. (Note that delay results will be given in terms of transmission time for one ATM cell and results are, therefore, independent of the actual data rate on the radio channel).

Referring to Equation (13), since P_0 must be non-negative,

$$\begin{aligned} \gamma &\geq \frac{\overline{M}}{1+\overline{M}} \quad \text{, or} \\ q &\geq \left(\frac{\overline{M}}{1+\overline{M}} - \frac{J-N}{J}\right)\frac{J}{N}. \end{aligned} \tag{21}$$

In Expression (21), q represents the probability that a given remote does not generate an information request in a given frame. (Note that this is consistent with the earlier definition of q.) This probability is further assumed to be the same for all the remotes. The minimum value of P_0, (the probability that the base station queue is empty at any given time) in the steady state, is governed by the value of q (Equations (21) and (13)). For a fixed frame length J, the value of P_0 should be as low as possible to achieve the highest utilization efficiency. This in turn implies that the value of q should be as low as possible, i.e., we wish to generate information requests at a rate sufficiently large so as to keep the data slots filled. However, Expression (21) indicates that some minimum value of q is needed to maintain stability. If equality holds in Expression (21), then $P_0 = 0$, and, given the system parameters J, N, and $\overline{M}$, the highest utilization efficiency is achieved. However, the right hand side of Expression (21) is negative if $J/N > (1+\overline{M})$. In this case, equality cannot hold (since q, being a probability, cannot be negative); P_0 cannot be zero, and the minimum value of P_0 occurs when q is equal to zero. This means that, on average, some portion of the data segment is always unused.

On the other hand, if $J/N < (1+\overline{M})$, then right hand side of Expression (21) is positive. $P_0 = 0$ is achieved when equality holds in Expression (21). However, this refers to a case when the polling cycles occur more frequently than need be, resulting in excessive overhead and therefore the *maximum utilization efficiency* (which may be achieved by properly choosing the controllable system parameters) is still not achieved. Since $q > 0$ in this case, on average, there are always wasted polling slots in each polling segment.

When parameters are chosen to satisfy $q = 0$ and $P_0 = 0$, optimal utility of both the the data slots and polling slots is ensured, and the maximum utilization efficiency [5] is achieved. For these conditions to be satisfied, the following equation should hold,

$$J = N(1+\overline{M}). \tag{22}$$

In this case, each remote generates one request per frame, thus ensuring maximum utility of polling slots. The number of information requests per frame in this case would be N. In response to these N requests, there would be on average $N\overline{M}$ information reply packets added to the base station queue. Thus, to maximize utilization efficiency, the length of the data segment should be chosen to be $N + N\overline{M}$ so that, on average, a polling segment occurs just when the base station queue has no more pending requests (See Equation (22)).

In practice, optimality would be maintained by varying the number of data slots from frame to frame, depending upon the number of remotes in the cell and a statistical estimate of the average information reply message length. It would, in fact, be necessary to include a signaling section in the frame [10] so that the base station could re-assign polling slots as remotes enter and exit the cell.

In Figure 3, we plot the highest achievable utilization efficiency (for various reply message lengths) as the data segment size J varies. If $J/N < (1+\overline{M})$, it is seen that, for a given average message length $\overline{M}$ and a given number of remotes N, the highest utilization efficiency, corresponding to minimum P_0, monotonically increases with frame length J and its maximum occurs when $P_0 = 0$. But when $J/N > (1+\overline{M})$, the highest utilization efficiency increases as J decreases, and is maximized when $q = 0$, as explained in the earlier paragraphs. In both cases, the maximum utilization efficiency is seen to be achieved when Equation (22) is satisfied.

Thus, we can see that by appropriately choosing the data segment size, maximum utilization efficiency close to 95 % can be achieved.

In Figure 4, we plot the expected message delay versus utilization efficiency for various values of J, while keeping the N and $\overline{M}$ fixed. This figure also shows the results of simulation

[4] In addition an information request experiences an average delay of half a frame to reach the base station, after it is generated by a remote. This delay is ignored here.

[5] For any other choice of parameters, a better utilization efficiency *cannot* be achieved.

studies which are seen to match very well with the analytical approximations. It is interesting to note that in the regime of low utilization efficiency, the expected message delay is almost independent of the frame size. This may be attributed to the fact that, in this regime, an arriving information request would, with very high probability, get an information reply from the base station within the same frame and would not experience delay due to intermediate polling cycles or wasted data slots. It should be noted that the expected message delay refers to the time interval between the moment that a given remote's information request reaches the base station and the moment that this remote receives an information reply message from the base station. On average, an additional delay of $J/2$ slots will be encountered, representing the time taken for an information request to reach the base station, once it is generated by a remote; this additional delay was ignored in the above.

4.1 Admission Control

Equation (21) can be rewritten as the following

$$N \leq \frac{J}{p(1+\overline{M})} \tag{23}$$

This provides an admission control policy so as to maintain system stability. Let $\lambda = p/T$, where T is the total duration of the frame in seconds. Then, λ would represent the maximum rate at which remotes can generate requests, given the constraint that they can request only once per frame. By appropriate normalization, λ can be represented as a scaled version of $\lambda_0 = p/J$. Thus, for a given arrival rate, one can administer an admission control policy so as to ensure that QoS guarantees of mean delay bounds are maintained.

For a given value of λ, the values of p and T affect the burstiness of the traffic arriving at the base station queue. Hence, the mean delay could be different for different p and T. We administer an admission control policy based on the assumption that p and T are chosen so that the best delay performance is achieved for a particular λ. In Figure 5, the number of users per cell who can be admitted to a cell, for a given arrival rate λ, is plotted for various delay constraints. Based on the desired QoS guarantee on delay, this graph tells us the maximum number of users that can be admitted to the cell.

5 Modifying the protocol to accommodate real-time traffic

In the previous discussions we assumed the presence of variable bit rate non-real time traffic applications which include telnet sessions, web browsing sessions etc. This protocol may be modified to include both *real time* constant bit rate (CBR) and variable bit rate (VBR) traffic. The CBR traffic would consist of periodic bursts of data of constant size. The real-time VBR session would produce periodic bursts of data but the size of each burst would be different and the burst size would depend upon the traffic shaping imposed at the source remote. The data segment of the frame may then be partitioned into two portions, a portion dedicated for bursty non real-time traffic and the second portion in which the real-time traffic has priority. This is akin to the hybrid circuit/packet switched multiple access schemes proposed in many papers for different multiple access networks [13]. The boundary between the real-time and non-real time sections is movable in the sense that if there are a small number of real-time connections in progress, the unused capacity of the real-time portion of the frame may be temporarily be used by non real-time traffic. This movable boundry concept is depicted in Figure 6.

When a real-time CBR connection is established [6], a constant number of requisite time slots are dedicated for that connection. In each frame, the application would include a field to indicate whether the call is continuing to the next frame or whether it terminates in the current frame. This is depicted in Figure 7.

If the connection is a real-time VBR connection, each VBR connection is allocated a deterministic number of slots, as requested by the connection at the set up phase. This number might be equal to the average burst size or might be larger than the average burst size (This is determined by the policy and cost.). It is expected that the real-time VBR traffic is shaped by using either a leaky bucket or a token bucket scheme [14] to constrain the burstiness to within certain limits [7]. In each burst, a field is included to indicate the number of slots required for the next burst (Figure 7). If the aggregate number of slots requested by all the bursting connections, in a particular frame, is less than or equal to the number of slots in the real-time portion of the data segment, each connection is allocated the number of slots it requested in the immediately following frame. If the aggregate number of requested slots of all bursting connections is greater than the number of slots in the real-time portion of the data segment, then the scheduler would use a weighted fair allocation policy for allocation of slots in the next frame. It would first identify those connections whose current burst size is less than or equal to their average burst size and allocate to these connections, the number of slots each of these connections requested. For the remaining connections whose burst sizes are larger than their average burst size, the scheduler would first allocate to each connection, a capacity equal to the average burst size of that connection. The remaining capacity (if any) in the real-time portion of the frame is then shared among these *connections with excessive burst sizes* by using a weighted fair allocation methodology by which each of the connections receives a share proportionate to its current burst size. Excessive packets might be dropped or an attempt may be made to reschedule them in a subssequent frame. It is to be noted that the base station might use a dedicated spread spectrum slot in order to indicate to each real-time connection, the slots allocated to it. This would be a broadcast transmission destined for all remotes within the cell. Furthermore, an admission controller which would reject new real-time connections if enough capacity is not available, is needed.

It is important to note that regardless of whether the connection is a real-time CBR or a real-time VBR connection, the communicating remote must send a pilot tone to the base station prior to every burst to enable the base station to compute the appropriate set of weighting co-efficients. The base station will therefore have to indicate to each communicating remote, the time at which the remote has to send the pilot tone. As mentioned earlier, a seperate broadcast packet might be transmitted by the base station (with robust spread spectrum coding) to relay this information to every remote.

6 Conclusions

In this paper, we proposed a polling-based medium access protocol useful for fixed point, pedestrian, and vehicular cellular ATM networks. Its primary virtue is its compatibility with the use of a smart array antenna as needed to abate multipath and adjacent cell interference. It was shown that by appropriately choosing the

[6] Note that this connection might be initiated in either the base to remote or the remote to base direction.

[7] If no constraints are imposed on the burstiness of the incoming connections, a lot of packets may be dropped due to periods of congestion due to simultaneous large bursts of connections.

data segment length, a utilization efficiency of close to 95 % may be obtained. The optimal length of the data segment of the frame, given the number of remotes in the cluster and the average message length, can be determined by the base station which can *adapt* itself to varying traffic statistics to achieve this high utilization efficiency. Although we considered only a request/response kind of system wherein users retrieve large files from databases in the wireline network, it could be modified to enable arbitrary traffic models, including constant and variable bit rate traffic classes. For constant bit rate traffic, data slots may be reserved appropriately at periodic intervals. For variable length remote messages, the current burst could inform the base station of the number of slots needed for the next burst, and the base station could appropriately allocate slots at the appropriate time.

As mentioned above, a signaling field would, in general, be needed to admit users to a cell and to accommodate hand-off. During this signaling period, any remote which wishes to hand-off to the base station would put in its request. Depending on the frequency of hand-offs, the size of the signaling field could be appropriately chosen. The effect of including such a field on the performance of the proposed scheme is a topic in need of further attention.

References

[1] A.S.Acampora and J.H.Winters, “A Wireless Network for Wide-band Indoor Communications”, *IEEE JSAC*, pp. 796-805, vol. SAC-5,No.5, June 1987.

[2] A.S.Acampora and J.H.Winters, “ System Applications for Wireless Indoor Communications”, *IEEE Comm. Mag*, pp.11-20, vol.25, No.8, August 1987.

[3] M.Naghshineh *et al*, “ Issues in Wireless Access Broadband Networks”, *Fifth WINLAB Workshop on Third Generation Wireless Networks,* April 1995.

[4] M.J.Gans *et al*, “ High Data Rate Indoor Wireless Communications Using Antenna Arrays”, *in proceedings of PIMRC’95*, pp. 1040-1046, 1995.

[5] W.C.Jakes, Jr., Ed., *Microwave Mobile Communications*, Wiley,NY, 1974.

[6] *IEEE Personal Communications Magazine*, Special issue on Wireless ATM, August 1996.

[7] Z.Zhang and A.S.Acampora , “Performance of a Modified Polling Strategy for Broadband Wireless LANs in a Harsh Fading Environment”, *in proc GLOBECOM’91.*

[8] R.Kohno *et al*, “Spread Spectrum Access Methods for Wireless Communications”, *IEEE Comm. Mag*, pp.58-67, January 1995.

[9] S.Glisic and B.Vucetic, *Spead Spectrum CDMA Systems for Wireless Communications*, Artech House Publishers, 1997.

[10] A.S.Acampora and M.Naghshineh, “ Control and Quality-of-Service Provisioning in High-Speed Microcellular Networks”, *IEEE Personal Communications Magazine*, pp. 36-43, vol.1, No.2, Second Quarter 1994.

[11] S.M.Ross, *Applied Probability Models with Optimization Applications*, Dover Publications Inc., NY, 1992.

[12] J.F.Hayes, *Modelling and Analysis of Computer Communication Networks*, Plenum Press, 1984.

[13] W.Lidinsky and D.Vlack, *Perspectives on Packetized Voice and Data Communications*, IEEE Press, 1990.

[14] P.Ferguson and G.Hutson, *Quality of Service: Delivering QoS on the Internet and in Corporate Networks*, John Wiley and Sons, Inc. 1998.

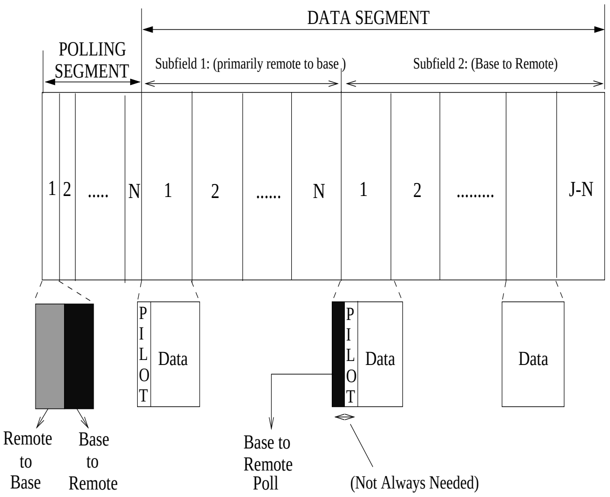


Figure 2: Basic Timing Diagram for Proposed Medium Access Protocol

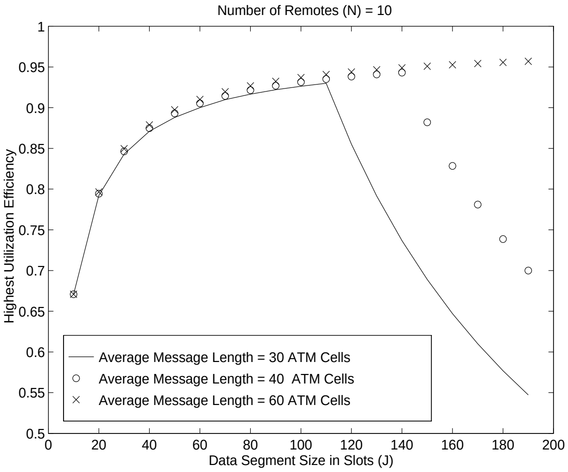


Figure 3: Utilization Efficiency vs Data segment size for varying message lengths

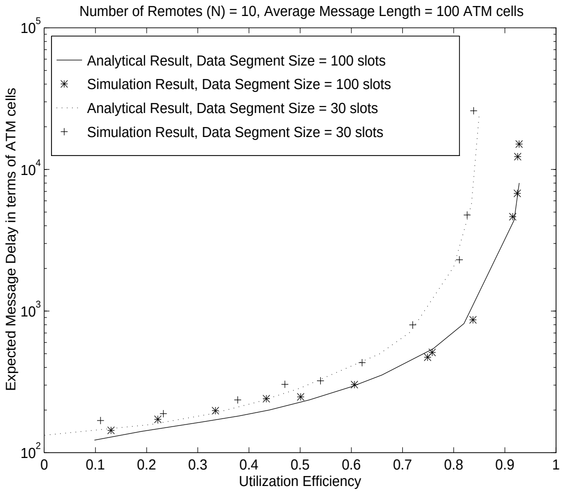


Figure 4: Expected Message Delay vs Link Throughput for N=10, J=30 & 100

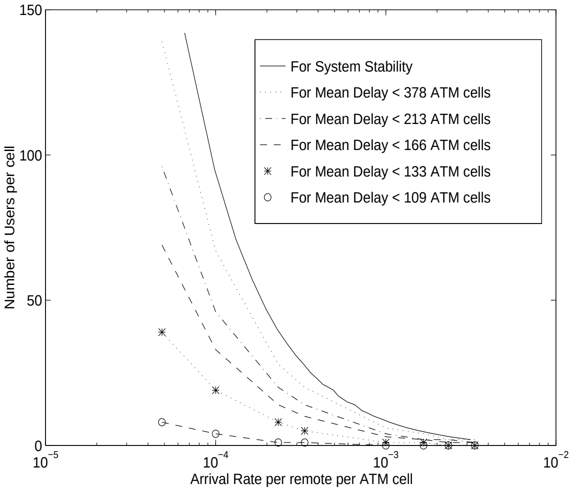


Figure 5: Mean Message Delay vs Expected Number of Users

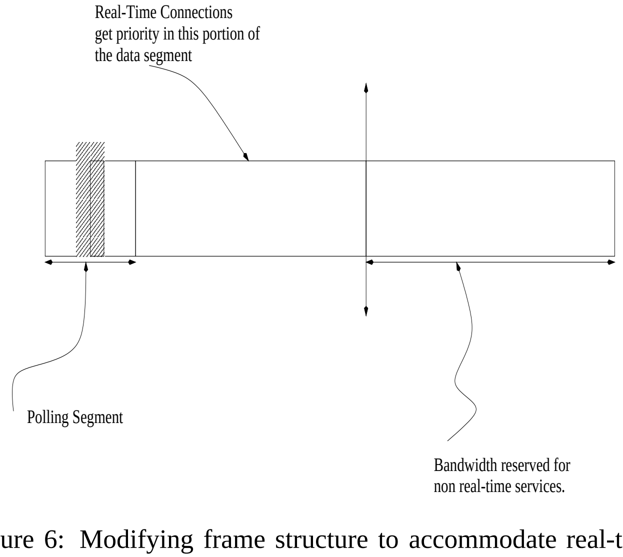

Real-Time Connections get priority in this portion of the data segment
Polling Segment
Bandwidth reserved for non real-time services.

Figure 6: Modifying frame structure to accommodate real-time connections

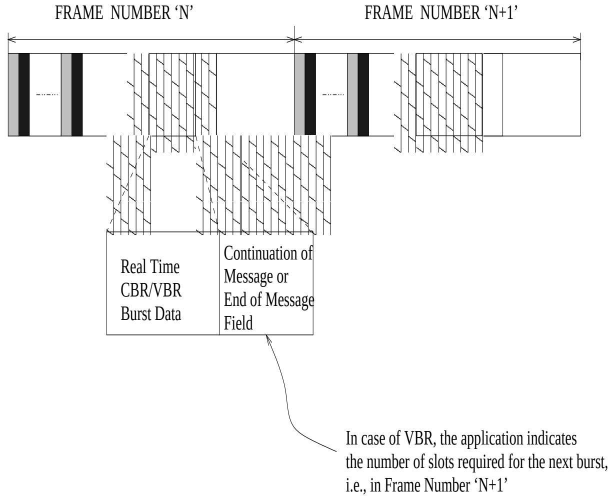

FRAME NUMBER 'N'
FRAME NUMBER 'N+1'
Real Time
CBR/VBR
Burst Data
Continuation of
Message or
End of Message
Field
In case of VBR, the application indicates
the number of slots required for the next burst,
i.e., in Frame Number 'N+1'

Figure 7: In band signaling for real-time connections