DUAL RATE MULTIPLE ACCESS
A RADIO LINK PROTOCOL

BY SIVARAMAKRISHNA VEEREPALLI

A thesis submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Electrical Engineering

Written under the direction of

Professor Roy D. Yates

and approved by

____

____

____

New Brunswick, New Jersey

October, 1996

ABSTRACT OF THE THESIS

Dual Rate Multiple Access
A Radio link Protocol

by Sivaramakrishna Veerepalli, M.S.

Thesis Director: Professor Roy D. Yates

In the recent past, there has been an increased demand for Cellular/Personal Communication Services. However, the increase in demand has not been matched by an increase in the available spectrum. This has resulted in an urgent need for the development of efficient techniques for spectral utilization. Efficient utilization of the spectrum could be achieved through improvements in various modules of a cellular system. This thesis suggests a method for improving spectral efficiency through the design of an efficient Media Access Control (MAC) Protocol.

In this thesis, we propose and analyze the Dual Rate Multiple Access (DRMA) protocol. The DRMA system supports high and low packet rates for voice users. High rate service is offered to active users with talkspurts, while the low rate service is used to poll inactive users. The polling strategy for inactive voice users enables statistical multiplexing of voice talkspurts. The low rate service is also used to generate a low bit rate representation of the background noise and could also be used as an embedded full time control channel for the voice users.

We also examine the integration of data and voice in the DRMA system. The residual bandwidth that goes unused by voice is assigned to the data users. Note that explicit priority is given to the voice users. An appropriate media access strategy for

data users is dependent on the type of data messages. In this work, we integrate a random access protocol for an email messaging data service. We find through simulation and analysis that DRMA affords a considerable increase over TDMA in spectral utilization in terms of the number of simultaneous calls that can be carried, and also in terms of the bandwidth available for data messaging.

Acknowledgements

I would like to thank Professor Roy D. Yates for sharing his knowledge and expertise, for being accessible and openminded in discussing my ideas. I greatly appreciate him for being a constant source of encouragement for me.

I also thank Professors David J. Goodman and C. Rose for participating in my thesis committee. I thank all the students at WINLAB for all those stimulating discussions.

Finally, I would like to thank Kishore Talluri for the help I received with programming and my parents for the constant moral support.

Dedication

Dedicated with love to my parents.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ARQ Automatic Repeat reQuest

BS Base Station

BW Bandwidth

CCRA Capetanakis Collision Resolution Algorithm

C-PRMA Centralized Packet Reservation Algoritm

CRA Collision Resolution Algoritm

DMV Dynamically Managed Voice system

DRMA Dual Rate Multiple Access

EIR Equipment Identity Register

HLR Home Location Register

HR High Rate

LR Low Rate

MAC Media Access Control

MS Mobile station

NW Network

PCN Personal Communication Network services

PCS Personal Communication Services

PRMA Packet Reservation Multiple Access

R-ALOHA Reservation ALOHA

RV Random Variable

S.D Standard Deviation

TDMA Time Division Multiple Access

VLR Visitor Location Register

Chapter 1

Introduction

Mobile radio systems play an important role within the commercial, business and public safety sectors in almost all industrialized parts of the world. Frequency allocations exist for communication between mobiles and their bases, from railway trains to the normal subscriber telephone networks, and for radio paging. There is sustained and substantial growth in every country where wireless systems have been introduced. Mobile radio usage continues to expand at the rate of 10% + per annum and is increasing [16]. This phenomenal growth in the past few decades and the increase in demand for capacity has resulted in a need for more efficient utilization of the scarce radio spectrum. To this end, improvements have to be made in the radio technology used in order to meet the capacity demands. In this work we present a method to improve spectrum utilization in TDMA systems.

Capacity improvements could be obtained through development in various modules of a mobile system. Improvements in the physical layer technology can offer a considerable capacity enhancement. However, for a given physical layer, development of efficient protocols also can lead to improvements in efficiency.

This thesis proposes a protocol to gain improvements in bandwidth efficiency and at the same time to solve certain inherent problems present in earlier TDMA protocols.

1.1 Wireless Network Architecture

Three network elements are essential to provide wireless access to information networks: wireless terminals, base stations and switches. A base station is the fixed end of the radio channel that links wireless terminals to the wire line information networks. The switch provides the interface between the wireless and fixed networks. The switch also controls communication within the wireless network infrastructure, a collection of the base stations, switches and supporting hardware and software [17].

The wireless service coverage area is divided into smaller units called *cells*. Each cell consists of a base station (BS) as the central node that serves a collection of mobile stations. Each mobile station (MS) communicates with the BS and vice versa through wireless links (shown as dotted lines in the figure 1.1). The cell is allocated a fixed or variable amount of transmission bandwidth in the form of discrete time or frequency channels. The base stations communicate with the mobile stations through *downlink* channels while the mobiles transmit to the base station through the *uplink* channels.

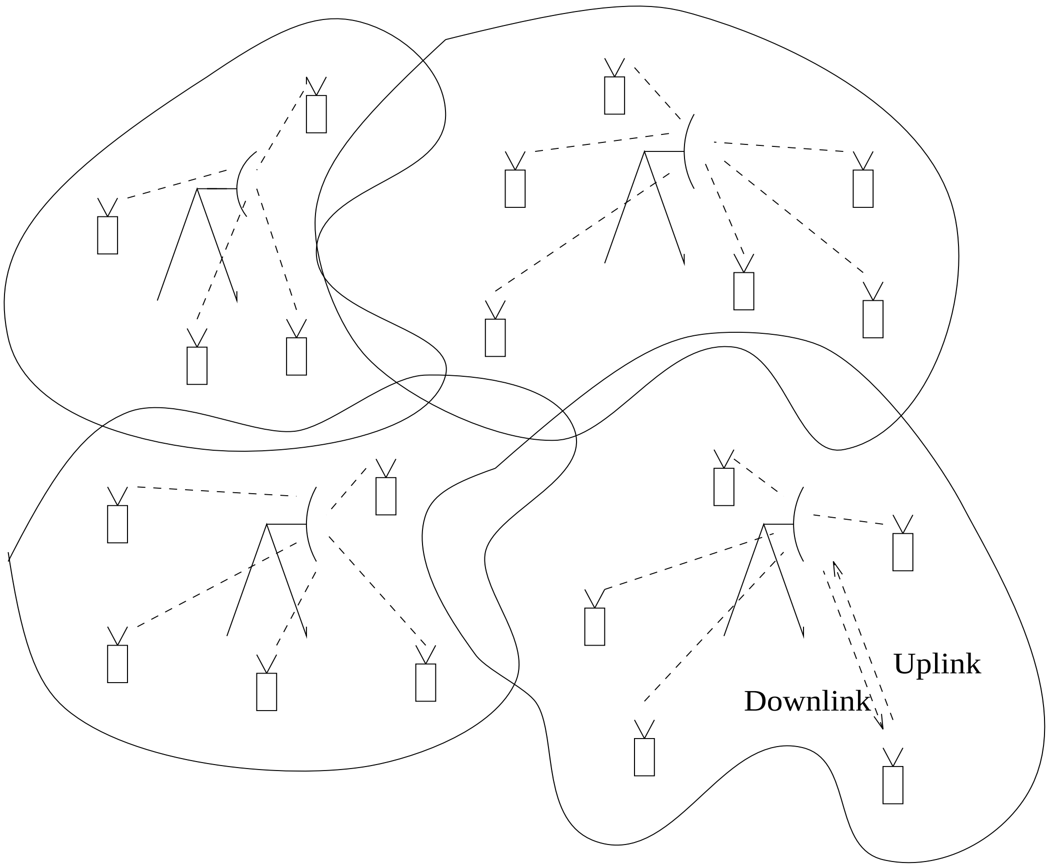

Figure 1.1: Wireless Network Architecture.

Base stations are linked to the switch through a wire, fiber or microwave links. Figure 1.1 shows a very simple wireless network.

1.2 Bandwidth Efficiency

The introduction of Personal Communication Services (PCS) has brought with it a host of new possibilities for communication. Apart from voice service, PCS also implies the need for data services and possibly video. Access to databases, e-mail, voice-mail and fax should become available to the voice user. These services should be feasible at a low cost.

In order to meet the demand for capacity, a larger number of smaller cell sites could be implemented resulting in greater channel reuse. However, this results in higher equipment costs. Otherwise, the number of users served by a fixed amount of bandwidth would have to be increased, i.e., improvements in spectral efficiency for a given physical layer technology are required in order to provide the service at reasonable cost and at the same time meet the demand. A few approaches to improving the spectrum efficiency include [7]

- higher level modulation
- lower bit-rate, robust voice coding and modulation that requires less channel coding
- less overhead
- better cochannel interference performance
- better adjacent-channel interference performance
- flexible bandwidth through statistical multiplexing

Here we deal with the last of the above mentioned solutions. Due to the reasons explained in the following sections, circuit switching is inherently inefficient, or, more accurately, it leads to inefficient use of the spectrum by not allowing the users to adjust their bandwidth to meet their actual communication needs. In contrast, a packet switched system can allocate a user as little bandwidth as is needed and therefore can be more efficient. Experience has shown that digital cellular systems are capable of implementing flexible bandwidth architectures with relative ease. TDMA systems are probably at an advantage here [17], [16], [7].

1.3 Statistical Multiplexing

The integration of voice and data is a requirement for the future communications systems. In an ordinary TDMA system when a channel is assigned to the user to establish a call, the channel is dedicated to that call for the entire duration of the call [2]. Thus if there were N slots in a TDMA frame, the maximum number of calls that could be simultaneously supported by the system would be N. Typically, neither voice nor data users require the full capacity of the link continuously. Periods of activity alternate with idle periods. Thus, assigning bandwidth on demand, instead of a circuit switched dedication of a fixed amount of bandwidth, increases channel utilization. Integrating voice and data services on the same communications link further improves channel utilization.

Packetized speech consists of active periods called *talk spurts* and inactive periods called *silent gaps*. The length of the talkspurt and the silent gap distributions depend on the type of voice activity detector used. The fraction of time a user is in the talking state is known as the *voice activity factor*. Typical values of the voice activity factor are in the range of 0.35 to 0.50 [13].

1.4 Motivation

Protocols which exploit speech activity detection in order to improve bandwidth utilization are generally known as Dynamically Managed Voice (DMV) Systems [13]. DMV systems utilize speech activity detection to exploit speech idle time and variable bit rate coding to exploit non-stationary speech statistics. The emphasis here is on systems using speech detection. DMV systems also utilize variable speech burst delay and speech clipping.

Prior work has been done on DMV protocols that utilize statistical multiplexing in order to gain bandwidth efficiency [11]. These protocols include reservation ALOHA and Packet Reservation Multiple Access (PRMA). Extensive work has been done on PRMA [8], [11], [12], [19].

PRMA utilizes a random access mechanism for the users to contend and gain access to the channel whenever the user has a talkspurt to transmit, similar to the ALOHA protocol. For PRMA, voice is packetized as 1 packet per frame i.e., once the user gains access to a channel (which is a time slot in a TDMA frame), the channel is reserved for the user until the end of the talkspurt, during which time he transmits 1 packet per frame. At the end of the talkspurt, the user relinquishes control of the channel. This process continues until the end of the call and thus more than N simultaneous calls can be supported by a system which contains N slots per frame.

PRMA has certain inherent disadvantages. When the user releases the channel at the end of the talkspurt, there is no channel between the two end users. This results in a *dead line* between the two users. A connection that appears to the listener as a *dead line* during the silence periods is perceived as inferior [21], [22]. Consequently it may be desirable or even necessary to include a simulated background noise, a *comfort noise* that makes the line sound alive between talkspurts. It has been found that such approximations cannot be made to sound too realistic and are not too comforting to the listener [21], [22]. Alternatively, a multiaccess protocol that supports a low bit rate approximation of the actual background noise during silence intervals may be preferable.

In addition, it would be desirable to have a full time control channel in which it would be possible to send or receive messages regarding power control or handoff. It should be apparent that such control channel would not need a very high bit rate. In the basic PRMA protocol [8], a user has no accessible control channel during the silence period. Although it is certainly possible to add a control channel on a separate band to the PRMA system, it would be more desirable to have a single communication channel that multiplexes the voice conversations and provides a full time control channel.

In this report, we describe Dual Rate Multiple Access (DRMA), a DMV system that statistically multiplexes voice conversations and provides each user a secondary channel that can be used for control messages as well as for a low bit rate representation of the background noise. The essence of DRMA is that the system supports high and low packet rates. During a talkspurt, a user would normally use the high rate while the low rate is used during silence periods. Appropriate packet header formats under low and high rates would permit an embedded full time control channel. In addition, the low rate service is used to establish (i.e. make a reservation for) the high rate connection. This reservation process will be contention free. As we shall see, the low packet rate will be closely tied to the maximum permissible packet delay.

Statistical multiplexing in the DRMA context, employs the idea of multiplexing talkspurts of all voice users sharing the same communication bandwidth. This results in more than N simultaneous calls. There would still be some bandwidth left that cannot be used due to the voice protocol design constraints. This remaining bandwidth could be utilized by the data users.

Chapter 2

Dual Rate Multiple Access Description and Simulation

Dual Rate Multiple Access or DRMA is a media access control (MAC) protocol for the radio link. It is based on a polling mechanism (for the voice users) and utilizes statistical multiplexing with a talkspurt reservation strategy. This protocol multiplexes independent mobile stations onto a common radio resource and its performance is influenced by the voice activity factor. Two rates of service are supported, high rate and low rate. The high rate service is used for information transmission while the low rate is used either for control purposes and/or to eliminate the *dead line* problem.

2.1 Medium Access Control Protocols

All wireless networks or fixed networks consist of a collection of nodes that must share the network's transmission capacity. Some means of controlling access to the transmission medium is needed to provide for an orderly and efficient use of that capacity. This is the function of a medium access control (MAC) protocol [20]. A MAC layer can be viewed as being present in between the physical layer and the Data link layer in the seven layer OSI model. The implementation of a MAC is either centralized or distributed. In a centralized scheme, a controller is designated that has authority to grant access to the network. In a decentralized network, the mobile stations collectively perform a medium access function to dynamically determine the order in which the stations transmit.

The advantages of a centralized scheme include greater control over access for providing priorities, overrides, and **guaranteed** capacity and relatively simple access logic at each node. The disadvantage is that it has a single point of failure.

We can categorize MAC techniques, based on bandwidth allocation, as either inflexible or flexible. With inflexible techniques a fixed amount of bandwidth is dedicated to a connection. This is the approach used in circuit switching, FDM and synchronous TDM. Such systems waste slots due to voice inactivity. It is preferable to allocate capacity on demand, in a flexible manner. Three types of the latter category are: round robin, reservation and contention.

2.1.1 Round Robin

Here, each station in turn is given the opportunity to transmit. The node, may decline to transmit or may transmit subject to certain constraints. When the station is finished, it relinquishes its turn , and the right to transmit passes to the next station in logical sequence. Control of sequence may be centralized or distributed [20].

When many nodes have data to transmit over an extended period of time, round robin techniques can be very efficient. A slot is wasted whenever a node decides to pass the turn to the subsequent node, in order to inform the base station/other nodes regarding its status. If only a few nodes have data to transmit over an extended period of time, there is a considerable overhead (in the form of the wasted slot) in passing the turn from node to node, since most of the nodes will not transmit but simply pass their turns.

2.1.2 Reservation

This is a very suitable technique for information sources, either voice or data with stream traffic. Time on the medium is divided into slots like in synchronous TDM and a node wishing to transmit reserves future slots for its use. Reservations may be made in centralized or distributed fashion. Once the node has nothing more to transmit, it may relinquish the use of the channel.

2.1.3 Contention

All stations with information to transmit contend for access to the channels. These techniques are distributed in nature. The main advantage is that they are simple to implement and efficient under light to moderate load. For some of these techniques however, performance tends to collapse under heavy load [20]. The PRMA protocol makes use of contention with reservation [11].

2.2 The DRMA Protocol

When a user begins a talkspurt, it is necessary to buffer the speech burst until a channel becomes available. Generally, speech bursts transmitted in this manner encounter variable delay. A second method (known as speech clipping) is to discard speech packets when the transmission capacity is not available. A number of speech clipping strategies with different subjective effects exist. In reservation systems, delay constraints dictate front-end clipping. It has been found through experiments [13] that, front-end speech clipping is much more tolerable than either mid speech clipping or back-end burst clipping. A third method is to apply variable rate coding to the incoming speech, the rate being dependent on the traffic load on the DMV system [13].

In DRMA, we utilize all the above mentioned methods with minor modifications, such as, in variable delay, we make sure that we do not buffer packets for more than its maximum delay. If after the maximum delay, there is still no transmission capacity available, the packet is discarded. If the DRMA frame is L msec long and the maximum packet delay is D msec, where $L < D$, each packet could be buffered for a duration of $d = \frac{D}{L}$ frames. For example, if the DRMA frame consists of 11 slots each 1 msec long and the maximum packet delay is of the order of 30 msec, each packet can be buffered for a maximum time of 3 frames duration. Also, a dual rate of service is provided in order to accommodate users with two possible rates; the assigned rate may depend on the network load conditions. The dual rate of service provides additional advantages which are described below.

Consider a time slotted system that serves a collection of voice users. Each user is characterized by a two state, silent or active, continuous time Markov chain. The active state corresponds to being in a talkspurt. During a talkspurt, a user generates one packet every frame deterministically. Just as in PRMA, the N slots define a frame. When the user is silent, no packets are generated. Each packet can endure a delay equal to dN slots or d frames.

In DRMA, each inactive voice user is guaranteed a slot once every d frames while an active user is given a slot in every frame, as long as a slot is available. In the event that there are more active users than there are slots in a frame available for active users, some of the active users will be given the low rate service. This type of service can implemented in a straightforward manner. Consider a user that makes a transition from silent to active with the generation of the first packet of a talkspurt. This user will have a guaranteed slot available within d frames. Since a packet can endure a delay of up to d frames, the user can and will wait until his guaranteed slot to transmit the packet. In the packet header of that transmission, the user will request the high rate service to begin in the subsequent frame. While the user is active, the system will assign the user a slot to use in every frame as long as the slots are available. If a slot is not available, the active user is guaranteed a slot d frames later but may drop packets in the interim. Finally, we note that if a slot becomes available while a blocked active user is waiting for the guaranteed slot, this user will be allocated the slot. Effectively, when too many users are simultaneously active, those active users that are forced to use the low rate will drop $d-1$ out of every d packets while the system is overloaded. As we shall see, for reasonable values of packet delay d, the advantages afforded by DRMA can be obtained at a reasonable cost.

DRMA has both advantages and disadvantages relative to PRMA. First, we observe that if the number of users K is less than the number of slots N, then DRMA never drops a packet. By comparison, under PRMA with $K < N$, packets are occasionally dropped due to the contention process used to make talkspurt reservations. When $K > N$, both DRMA and PRMA will drop packets. However, it is unclear whether packet dropping under DRMA will be greater due to the slots guaranteed for the low rate users or lower since no contention occurs. We will investigate these issues in the following sections.

2.3 Simulation Model

In the simulations presented in this section we include only periodic information sources namely speech sources. A fixed number of active calls in the system are assumed for the duration of the simulation.

2.3.1 Speech Activity Model

Speech activity detectors can detect two types of talkspurts and silent periods: The principal spurts and gaps are related to talking, pausing, and listening patterns in a conversation. The "minispurts" and the "minigaps" are a result of short silent intervals that punctuate continuous speech.

In DRMA, we implement what is described as a slow speech activity detector [12]. The minigaps are not detected by the speech activity detector leading to a slight over estimation of the actual speech activity and a reduction in the number of speech/silent transitions. In the simulation model, the talkspurts have an exponentially distributed duration with a mean value t_1 of 1.0 sec. The silent spurts have an exponentially distributed duration with a mean value t_2 of 1.35 sec. This process can be modeled as a two state Markov process as shown in the Fig. 2.1.

The transition from active state to the inactive state, i.e, the probability that a talkspurt with mean duration t_1 ends in a time slot of duration t_{slot}, is given by

$$\gamma = 1 - \exp(-t_{slot}/t_1) \tag{2.1}$$

and the transition probability from the inactive or silent state to the active state , i.e. the probability that a silent period with mean t_2 ends in a time slot of duration t_{slot}, is given by

$$\sigma = 1 - \exp(-t_{slot}/t_2) \tag{2.2}$$

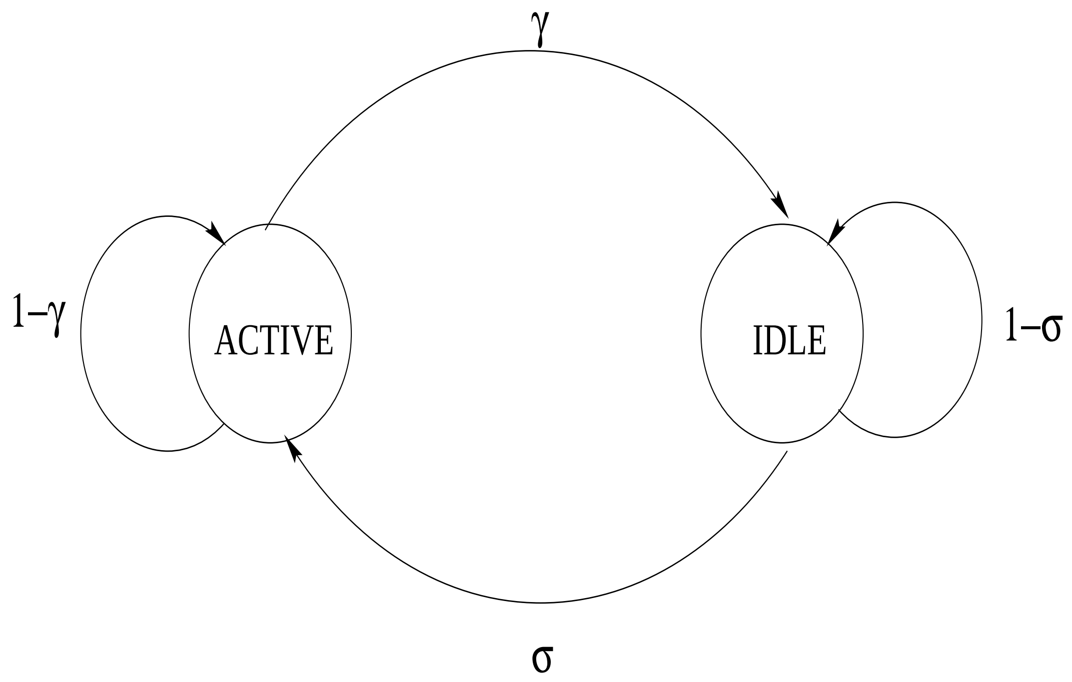

Figure 2.1: Two state Markov model of the Speech activity detector.

The probability that a user is in the active state is known as the voice activity factor given by the equation

$$\text{Voice Activity} = p = \frac{t_1}{t_1 + t_2} \tag{2.3}$$

For the parameters given above we find that the voice activity is equal to 0.426. In other words, a terminal is assumed to have packets to transmit in 42.6% of the frames.

2.3.2 Transmission Channel Model

An error free transmission channel is assumed for the simulation model. Thus, once an active user gains access to high rate slots, that user always transmits packets to the base station without any errors or any packets being dropped. This is of course an optimistic assumption. In a real environment, however, channel impairments due to radio propagation effects such as noise, shadowing, fading etc. would result in additional errors in the transmitted signal. MAC protocols e.g., Centralized PRMA (C-PRMA), that use retransmission of voice packets destroyed due to channel conditions has been considered in [3]. C-PRMA employs retransmission of the speech packets and it depends on the frame duration [3]. For data communications, any one of the transmission

Voice Users	Packets Dropped	Total Packets Generated	Dropping ratio
14	4001	1725481	0.0023
15	11068	1829302	0.0060
16	32053	1947707	0.0164
17	64506	2028187	0.0318
18	113398	2115195	0.0536
19	198692	2140573	0.0928
20	243091	2225132	0.1092
21	401583	2185301	0.1837

Table 2.1: Voice user packet dropping.

schemes such as Go-Back-N ARQ or selective repeat ARQ could be implemented.

2.3.3 Warm-up Period

At the start of the simulation, all voice users are assumed to have a talk spurt. To get the initial system state to a more typical state, a warm up period is introduced during which no statistical data is collected. Performance data is collected only after the expiration of the warm up period.

In the voice simulation, the warm up period is for the duration of 1000 frames, i.e. 11 sec as compared to an average talk/silence cycle of $t_1 + t_2 = 2.35$ sec.

The actual data is collected over a simulation period of 500,000 frames. The number of packets dropped during this period for a variable number of voice users is shown in Table 2.1.

2.4 Protocol Simulation

We shall now describe the design and the implementation of the DRMA frame structure. The voice users are classified into two types, Active (has a talkspurt) and Inactive (no talkspurt). The DRMA protocol will partition the frame into low rate (LR) and high rate (HR) slots. The LR slots will be used only by inactive low rate users. The HR slots are generally used for active users although they may be used by inactive users when there are very few active users. Users in the HR slots, when active, transmit packets in every frame until the end of their talkspurt. If there are three inactive users allotted to a low rate slot (say users x, y & z), then, if user x gets a guaranteed slot (G-slot) in an arbitrary frame 't', user y gets the slot in frame '$t+1$' and user z in frame '$t+2$' msec. User x gets to transmit again in frame '$t+d$'. Active users in the LR slots get to transmit packets in their guaranteed slots till they are assigned a HR slot. These slot requirements are depicted in Table 2.3.

Each user becomes active or inactive as determined by the speech model. When a user becomes active, it requests the high rate service. When such a request is made, the base station tries to swap the positions of an inactive user in a HR-Slot with the service requesting active user in the LR-Slot. However, there may be users who are active, but are not assigned a high rate slot as the number of active users is greater than the number of high rate slots. These active users are assigned a low rate slot.

A user who is denied a high rate request using a G-slot may be offered a reserved slot before making a subsequent request d frames later in the next guaranteed slot. It should be apparent to the reader that the base station can keep track of the bookkeeping in order to grant such reservations. An example of this bookkeeping is provided in subsection 2.4.1.

Let a set of d frames be known as a superframe. Say, there are K users in the system. Each low rate user is guaranteed one slot every superframe, while a high rate user gets one slot every frame within the superframe. This limits both the total number of users the system can carry as well as the number of simultaneous high rate users. Assuming that the maximum number of high rate users is an integer M, then each of these M users occupy one slot each in a frame. Thus $K-M$ low rate users must be accomodated in the remaining $(N-M)d$ slots of the superframe, where N is the total number of slots per frame. Therefore we see that the maximum number of high rate users that can be carried in a superframe will be the largest integer M such that

$$K - M \leq (N-M)d \tag{2.4}$$

From equation 2.4, we can conclude that

$$\begin{aligned} K - N &\leq (N-M)(d-1) && (2.5) \\ \left\lceil \frac{K-N}{d-1} \right\rceil &\leq N - M && (2.6) \end{aligned}$$

$$M \leq N - \left\lceil \frac{K-N}{d-1} \right\rceil \qquad d > 1 \tag{2.7}$$

Since equation 2.7 represents the maximum number of high rate users that could be accomodated without packet dropping, the maximum number of high rate slots in a frame is equal to M and therefore the number of low rate slots would then be equal to $N - M = \left\lceil \frac{K-N}{d-1} \right\rceil$. As can be seen from the equation above, the number of low rate slots in a frame is independent of the number of high rate users and hence the number of low rate users. It only depends on the total number of users and the number of slots per superframe. Each of these $\left\lceil \frac{K-N}{d-1} \right\rceil$ low rate slots (LR-Slots) supports more than one user since $[K - N] + \lceil \frac{K-N}{d-1} \rceil$ users are distributed over all the LR-Slots, while the high rate slots (HR-Slots) consists of only one user each.

In short, we see that DRMA can support M high rate users while providing a low rate service for the remaining $N - M$ users. By comparison, PRMA can sustain N high rate users but provides no low rate service. Although the ability of DRMA to support both a background noise process as well as a full time control channel for each user is a clear advantage, we will compare PRMA and DRMA only on the basis of actual voice packet dropping. Lastly, we observe that the analysis also is applicable when the users represent bursty data sources. In the active state, such a user would be transmitting a packet burst. In this case, much larger values of d may be possible because data users are less delay sensitive and would not require the background noise process.

2.4.1 Shift in Guaranteed Slot Sequence - An Example

An explanation of how the simulation works is given below with the help of an example for a clear understanding of simulation model. For purposes of explanation, consider that the frame is an 11 msec, 3 slot structure. Also assume that there are 4 users in the system.

The letters G, I and A (see Table 2.2) describe a guaranteed slot in an inactive period, an inactive slot in an inactive period and an active slot in an active period respectively. Let Table 2.2 represent the talk/inactive spurts of each user over nine frames.

Frame#	*User Number*			
	1	2	3	4
1	G	I	I	A
2	I	I	I	I
3	I	G	G	I
4	G	A	A	G
5	I	A	A	I
6	I	A	A	I
7	G	A	A	G
8	I	I	A	I
9	I	I	A	I

Table 2.2: User Talk/Silence spurts at any arbitrary time t. I denotes an Inactive user without a slot, G denotes an inactive user with a guaranteed slot, and A denotes an active user.

In the system, the position of the guaranteed slots for the users may vary due to swapping. This results in fewer packets being dropped. Before going into the details, we give a brief description of what Table 2.3 represents. In Table 2.3, the first slot is the LR-slot and the remaining two are HR-slots.

When the base station is requested for high rate service by a user, the base station can only swap users if the inactive user's new G-slot in the low rate slot (LR-slot) ensures that the distance between his previous G-slot in the HR-slot and his current G-slot in the LR-slot is less than or equal to the maximum packet delay. If no high rate slot (HR-slot) is available, then the user keeps dropping packets till an HR-slot becomes available. However the user does continue to transmit a packet every d frames in his G-slot.

Now for the Table 2.2, consider the physical implementation/simulation. It can be seen from Table 2.2 that, when user 3 became active in frame three, user 3 in its guaranteed slot informs the base station that it has an active spurt and hence needs to be assigned to an HR-slot. In frame 4 however, the BS is not in a position to swap the idle user 4 with the active user 3 as user 4 has his guaranteed slot in the fourth frame (see Table 2.3). If swapping were to be done here, user 4 would then get his G-slot in the sixth frame only and this results in a delay of more than the maximum packet delay before which user 4 is polled by the BS. This is certainly not acceptable to user

Frame #	LR	HR	HR
Frame 1	1G	2I	4A
Frame 2		2I	4I
Frame 3	3G	2G	4I
Frame 4	1G	2A	4G
Frame 5		2A	3A
Frame 6	4G	2A	3A
Frame 7	1G	2A	3A
Frame 8		2I	3A
Frame 9	4G	2I	3A

Table 2.3: Superframes illustrating swapping resulting in change of guaranteed slot sequence for user 4.

Frame#	*User Number*			
	1	2	3	4
1	G	I	I	A
2	I	I	I	I
3	I	G	G	I
4	G	A	A	G
5	I	A	A	I
6	I	A	A	G
7	G	A	A	I
8	I	I	A	I

Table 2.4: Modified guarantee slot sequence for user 4.

4. Since no other slots are available, user 3 drops a packet in the fourth frame.

In the fourth frame user 4 informs the BS that he is still idle. Thus in frame five, the BS swaps user 3 with 4 resulting in a HR slot for the active user 3 and a shift in the G-slot sequence for user 4 (compare Table 2.4 with Table 2.2 for the sequence of user 4). It can be seen from Table 2.3 and Table 2.4 that there are no more drops after the fourth frame as the number of active users is less than the number of slots in a frame.

Comparing Table 2.4 with Table 2.2 reveals that there is one less drop in Table 2.4 due to the swapping of users, resulting in a lower percentage of dropping for the simulation model as compared with the model for analysis seen in chapter 3.

Analysis of DRMA is done on a frame and superframe basis in chapter 3 and the results obtained from the superframe analysis are found to agree with the results obtained from the simulation of the actual protocol.

2.5 DRMA Simulation Results

Shown in the figures below are packet dropping results obtained from simulations. The graph shown in Figure 2.2 is obtained from the system simulation in which one slot in every superframe is going free (no one uses it) whenever there are even number of users (greater than ten) in the system. An even number of users results in one low rate slot not being packed to capacity. For $d = 3$, each slot could accommodate 3 users. An even number of users in a 11 slot frame would result in one of the LR-slots having only 2 users. This would result in one slot going waste every superframe (ie., every three frames). This is reflected in the step like quality of Figure 2.2. It can be seen from simulation that DRMA supports 15 users with $d = 3$.

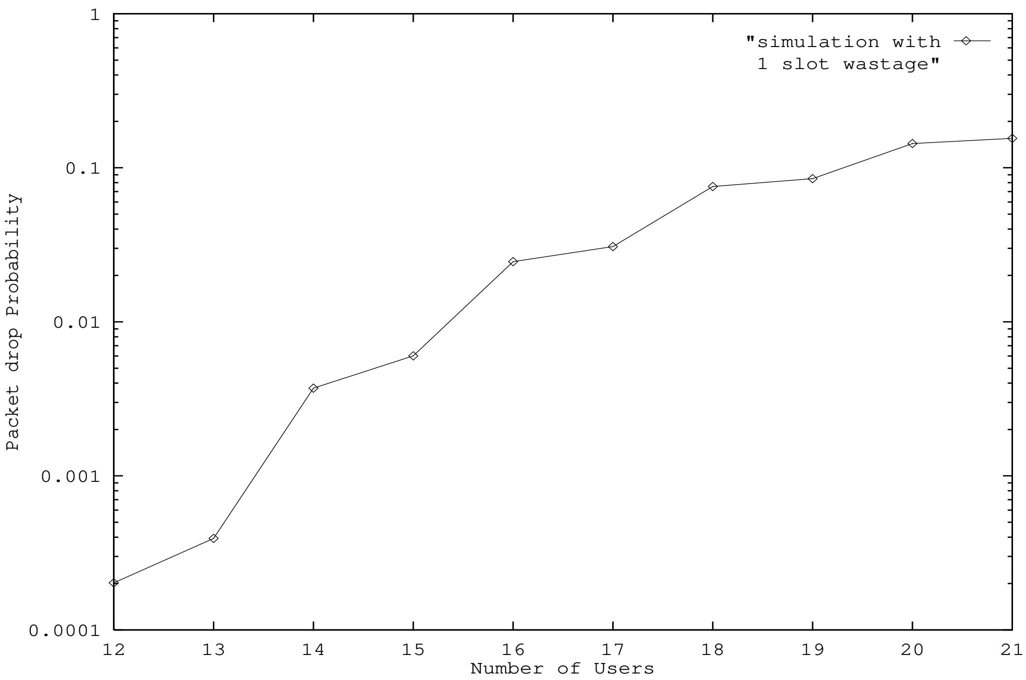

Figure 2.2: Protocol Simulation with wastage.

If it is assumed that the system can support an additional rate of transmission (besides the two already existing) in which, when there are even number of users greater than ten in the system, then, one active user in a low rate slot gets to transmit two packets every superframe. We say this system has "no wastage". The graph shown in Figure 2.3 is obtained through the simulation of a system which supports an additional transmission rate which is two thirds the full rate transmission.

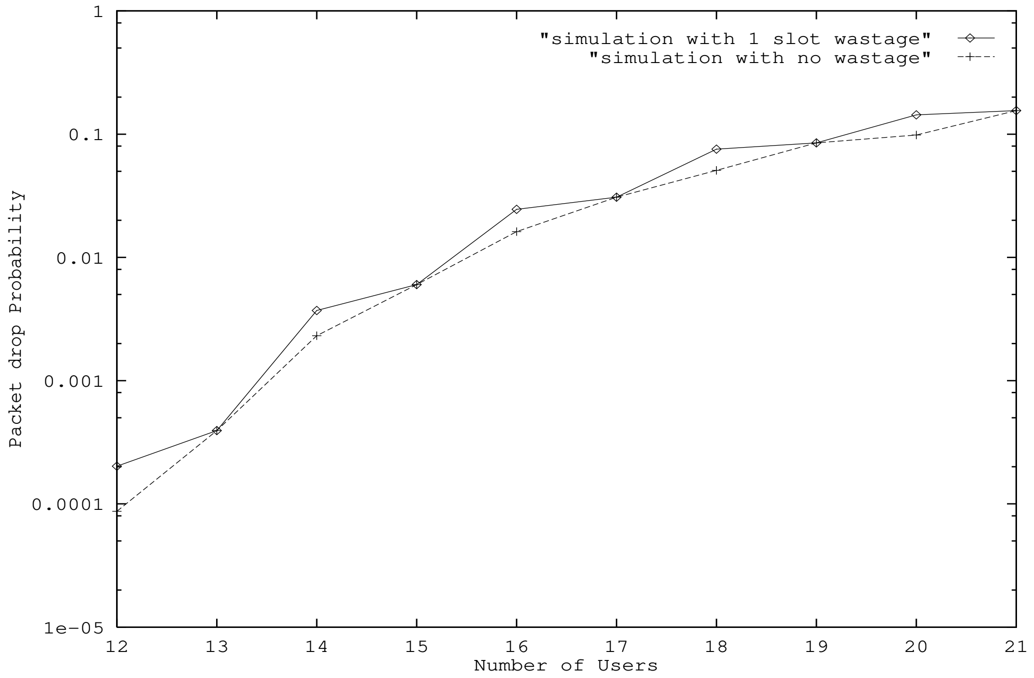

Figure 2.3: Simulation with and without additional rate of service.

It can be seen from the graphs in the Figure 2.3 that there is a reduction in the number of packets being dropped for an even number of users when an additional rate of service is offered.

The graphs shown above are from simulations in which the effect of buffering has not been taken into consideration. It should be noted that, by buffering, we mean that, the packets can be initially delayed for a period of d frames before getting dropped. In the unbuffered case, the packet is dropped as soon as the HR slot is not granted, whether or not the packet is d frames old. For a frame period of 11msec, the delay constraint is $d = 3$ frames. Introducing the buffering effect into the simulations yield the graphs shown in the Figure 2.4. It can be seen that the dropping percentage goes down further due to buffering effect. Also, with buffering, packet dropping percentage with 12 users is found to be almost similar to the dropping rate with 13 users wich

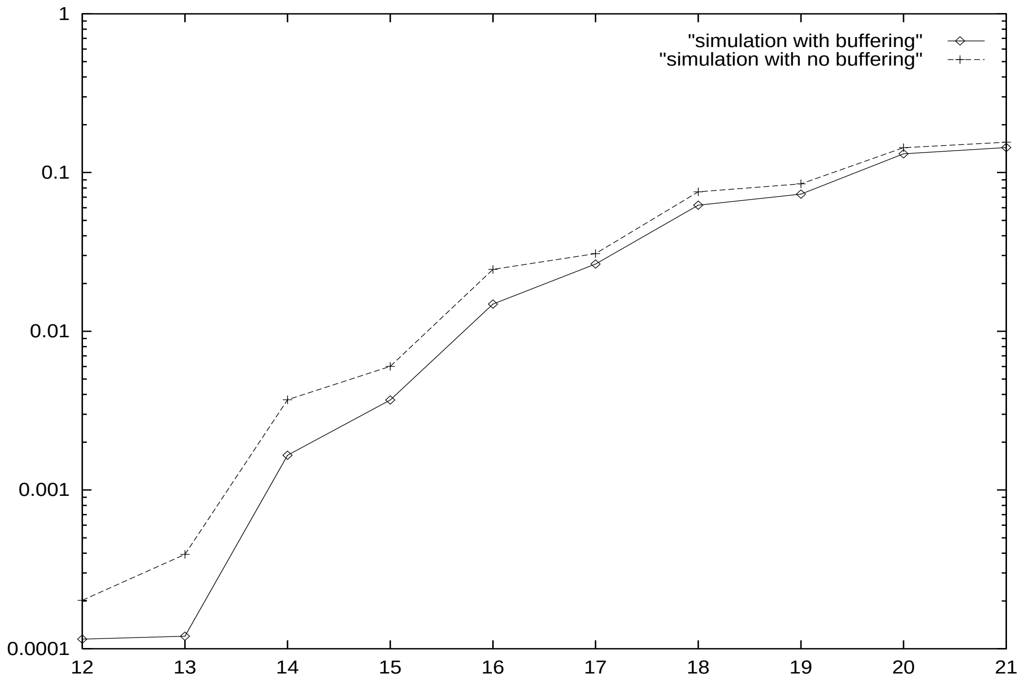

Figure 2.4: Protocol Simulations with wastage after buffering.

indicates that, when the number of users in the system is not too high, dropping is mainly dependent on the number of high rate slots available. It can be seen that as the number of users become very high, the difference in packet dropping between the two graphs becomes insignificant. This indicates that at higher load, dropping is affected mainly by the large amount of traffic in the system and buffering does not help to the extent that it does in the case of fewer number of users.

Chapter 3
Analysis of the DRMA Protocol

Approximate analysis of the Dual Rate Multiple Access protocol is performed at the frame level and at superframe level in the following sections and is followed by a comparison between the analysis and the simulation model. Frame analysis does not include the effect of the shift in the guaranteed slot sequence. It assumes that the guaranteed slot position remains fixed and therefore the model for frame analysis is somewhat pessimistic. Superframe analysis on the other hand, assumes only that packets must be sent within the superframe. It neglects the fixed position of the guaranteed slot and assumes that the low rate user is allotted one slot in every superframe. The superframe model is therefore optimistic and a study of these two models gives us a good estimate of the performance of the DRMA protocol.

3.1 DRMA Performance - Frame Level Analysis

Suppose the total number of users in the system is K. Each user has an independent packet arrival process. That is, a user will offer one packet every frame if active or one packet every d frames if inactive, independent of the packet dropping of the DRMA system. In an arbitrary frame, an arbitrary user with voice activity factor p will be active with probability p or idle with probability $1-p$. Hence, we would expect that an arbitrary user has a packet to send in an arbitrary frame with probability

$$b = p + (1-p)/d \tag{3.1}$$

In section 3.4 of this chapter, this conclusion is verified by modeling the two state source operating under the DRMA protocol by a Markov chain.

Consider an arbitrary frame i for which J_i users wish to send packets. Since each user has an independent packet arrival process such that in an arbitrary slot, each user requests to send a packet with probability b, J_i has the binomial distribution

$$P\{J_i = j\} = \binom{K}{j} b^j (1-b)^{K-j} \qquad 0 \leq j \leq K \tag{3.2}$$

Those packets corresponding to users with guaranteed slots will be transmitted. The remaining packets will be transmitted if enough slots are available. As long as the number of requests J_i does not exceed N, then all J_i requests can be honored; otherwise N packets are transmitted.

Given J_i, the number of packets dropped in frame i will be

$$D(J_i) = (J_i - N)^+ \tag{3.3}$$

Over L frames, the fraction of packets dropped over L frames will be

$$\frac{\sum_{i=1}^{L} D(J_i)}{\sum_{i=1}^{L} J_i} \tag{3.4}$$

Over a large number of frames, we have an average packet dropping probability

$$P\{D\} = \lim_{L\to\infty} \frac{\frac{1}{L}\sum_{i=1}^{L} D(J_i)}{\frac{1}{L}\sum_{i=1}^{L} J_i} = \frac{E[D(J)]}{E[J]} \tag{3.5}$$

where we have invoked the weak law of large numbers while overlooking the dependence of J_i and J_{i+1}.

Careful consideration will show this analysis to make a few simplifying assumptions. In particular, a user that is denied a high rate request using a guaranteed slot may be offered a reserved slot before making a subsequent request d frames later in the next guaranteed slot. In addition, a packet is offered only in one frame. That is, if a user is denied a high rate request in a frame, then the user will drop the next packet in the subsequent frame. This no buffering assumption is equivalent to assuming that on starting a talkspurt, the user's guaranteed slot occurs in the last possible frame. This is a somewhat pessimistic though not particularly significant assumption.

Figure 3.1 depicts the packet dropping in a DRMA system with an 11 slot frame as a function of the number of users. This is done for several values of the frame delay d. We see that most of the benefit of DRMA are obtained for fairly small values of

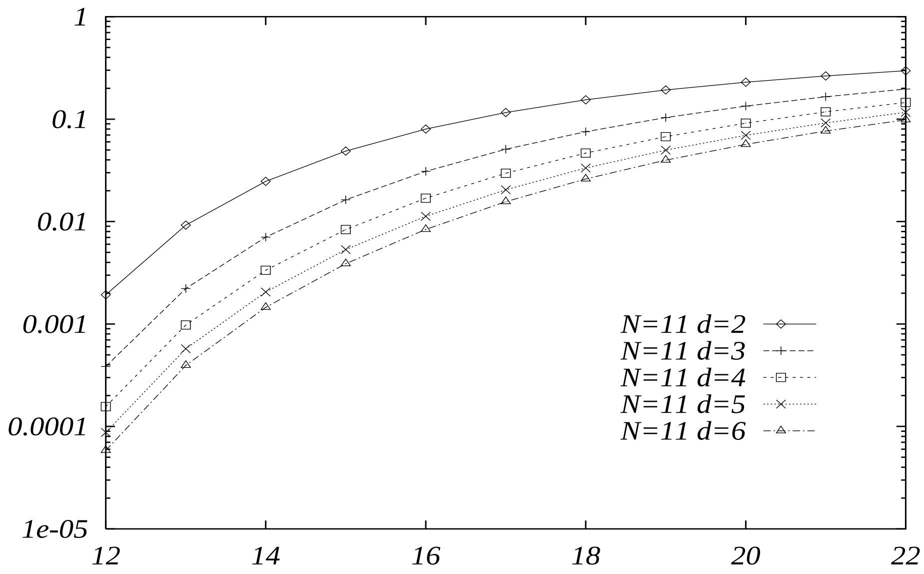

Figure 3.1: Frame Analysis: Packet Dropping vs. Number of Users in a system with a 11 slot frame.

the frame delay d. We see that at a 1% packet dropping rate, the DRMA system can support 14 users for $d = 3$ and 16 users for $d \geq 4$. For comparison, the basic PRMA scheme [11] supports 17 users at 1% packet dropping. Hence, in order to provide the dual rate service, the DRMA system supports three fewer users than the basic PRMA scheme.

3.2 DRMA Performance - Superframe Level Analysis

We have seen the analysis of DRMA performance earlier in section 3.1. That analysis was performed at a frame level, where each user is checked in every frame to see if he has a packet, either a packet from a talk spurt or a guaranteed packet generated by an Inactive user. It is assumed in the frame level analysis that each user gets a guaranteed slot once every d frames on a fixed basis starting from the frame in which the last talk spurt ended.

The analysis seen in this section is however performed on a superframe basis. Each superframe consists of d frames. It is assumed here that any user talk spurt length is in multiples of d frames. It is also assumed that every inactive user gets to transmit a guaranteed packet once in d frames, but not necessarily in frames which are multiples of d.

Suppose the total number of users in the system is K, where the frame length is equal to N slots. Then the expected number of users active in any superframe is given by the product of the number of users in the system and the probability that user is in the ON state in any arbitrary superframe.ie.,

$$E[S_A] = KP_{\text{on}} \tag{3.6}$$

P_{on} is the probability that a user is active in any arbitrary superframe. Let us now assume that a user is active on an average for J frames and inactive for L frames. Then,

$$\begin{aligned} \text{Superframes ON} &= \left\lceil \frac{J}{d} \right\rceil && (3.7) \\ \text{Superframes OFF} &= \left\lceil \frac{L}{d} \right\rceil && (3.8) \end{aligned}$$

Therefore,

$$P_{\text{on}} = \frac{E[\text{Superframes ON}]}{(E[\text{Superframes ON}] + E[\text{Superframes OFF}])} \tag{3.9}$$

P_{on} is found to be slightly greater than p, the voice activity factor. Consider an arbitrary superframe for which S_A users wish to send packets. S_A has the binomial distribution,

$$P\{S_A = s\} = \binom{K}{s} P_{\text{on}}^s (1 - P_{\text{on}})^{K-s} \tag{3.10}$$

Given that there are S_A active users and I inactive users in a superframe, the number of packets dropped per superframe is given by,

$$\begin{aligned} D(M) &= (dS_A + I - Nd)^+ \\ &= (dS_A + K - S_A - Nd)^+ \\ &= (S_A[d-1] + K - Nd)^+ && (3.11) \end{aligned}$$

It can be seen that d, K and N are constants. As long as the number of requests $(dS_A + I)$ do not exceed Nd, all requests can be honored; otherwise Nd packets are transmitted. Therefore the packet dropping probability is given by,

$$P\{D\} = \frac{E[(S_A(d-1) + K - Nd)^+]}{E[S_A(d-1) + K]} \tag{3.12}$$

The Figure 3.2 below depicts the packet dropping in a DRMA system with a 33 slot superframe(11 slot frame) as a function of the number of users. This is done for a packet delay constraint $d = 3$. We see that with $d = 3$ and a 1% packet dropping rate, the frame analysis indicates, the DRMA system can support 14 users while the more optimistic superframe analysis indicates that 15 users can be supported.

3.3 Simulation vs. Analysis

As has been stated in the previous chapter, it is found that the packet dropping percentage is smaller in the case of simulation than in the case of the analysis of the protocol at frame level. This can be seen from the Fig. 3.2.

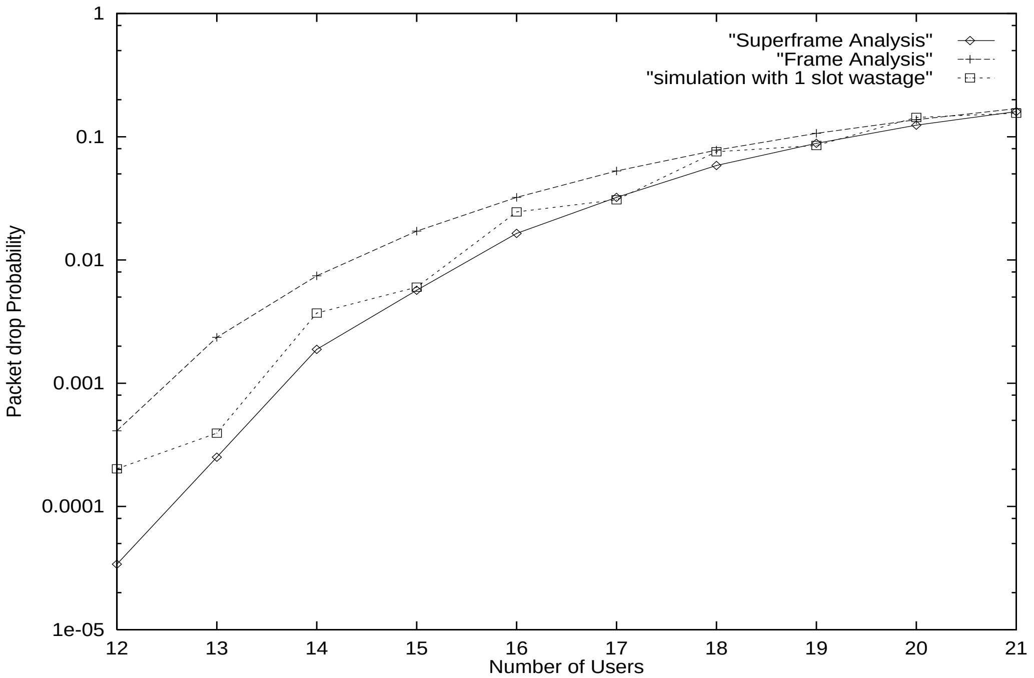

Figure 3.2: Superframe: Packet dropping as a function of the number of users with $d = 3$ and $N = 11$.

This is due to the fact that the analysis at the frame level assumes that an idle user has his guaranteed slot coming once every d frames, and that the position of this guaranteed slot is fixed throughout each inactive period. For example, see Table 2.2 the position of the guaranteed slots remain the same for all the following nine frames and the number of packets dropped in each frame would be equal to the difference between the number of active users in that frame and the number of slots per frame. In accordance with this, one packet is dropped in each of the frames 4 and 7 as the number of packets in both the frames is one greater than the number of slots in the frame. However, in the actual system, the position of the guaranteed slot for the user may vary due to swapping as illustrated in subsection 2.4.1. Swapping may result in the user getting his guaranteed slot before the end of d slots. This at the same time alters the guaranteed slot sequence for the user effecting the entire system, thus resulting in a lower dropping probability. This effect has been taken into account for simulation and yields a lower packet dropping probability. As has been mentioned in the previous chapter, any swapping is possible only within the superframe boundary, for the frame structure discussed earlier. The analysis that has been done at the superframe level takes into consideration that an idle user may get his guaranteed slot in any frame within the three frame superframe. This is similar to the actual implementation of the system where a user may get his guaranteed slot in any frame within the superframe. This has been verified by the results obtained from the superframe analysis. See Figures 3.2.

Figure 3.2 shows packet dropping as a function of number of users for both frame and superframe analysis. It is seen that dropping is less in the case of superframe analysis. It is also seen that the superframe analysis is closer to that of the DRMA simulation results.

3.4 DRMA Source Activity

We now derive equation 3.1. We observe that the packet arrival process of a DRMA user does not depend on the service the system provides those packets. A DRMA source is based on an underlying two state (OFF and ON) model of a voice source. When duration of the ON period is an exponential random variable A with mean $1/\mu$. In the ON state, the source generates a packet in every frame. While OFF, a packet is

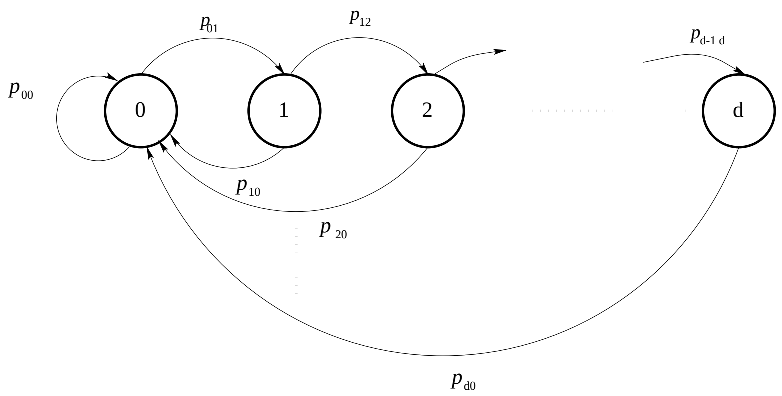

Figure 3.3: The states of a DRMA source.

generated every d frames. The duration of the OFF period is denoted by the exponential random variable S with mean $1/\lambda$. We use T to denote the duration of a frame.

We model this source by a Markov chain that describes a user's state at the beginning of each frame. Specifically, state 0 indicates that the user is in a talkspurt and wishes to transmit a packet in that frame. State i, $1 \leq i \leq d$, denotes the number of frames since the user last had a packet to transmit. When the user is in state d at the beginning of a frame, he is guaranteed to transmit a packet. If this packet is the first packet of a talkspurt, then the next state is 0; otherwise, the next state is 1 and the user waits another d slots until the occurrence of the next guaranteed slot. Note that after transmitting the first packet of a talkspurt in the guaranteed slot, the user enters state 0 even if the system denies the request for the high rate service. A sketch of the Markov chain appears in Figure 3.3. We now derive the transition probabilities shown in the figure.

Let $X(t)$ denote the state at the start of an arbitrary frame t. Suppose $X(t) = 0$ and the user has already transmitted $n-1$ packets of a talkspurt in frame t, the user transmits packet n and we can conclude that the duration of the ON period, A, satisfies $A > (n-1)T$. Moreover,

$$p_{00} = P\{X(t+1) = 0 | X(t) = 0\} \quad (3.13)$$

$$= P\{A > nT | A > (n-1)T\} \quad (3.14)$$

$$= e^{-\mu T} \quad (3.15)$$

The source enters state 1 when a talkspurt has been completed and the user begins a silence duration of length S. Note that if the source is in state i, $0 < i < d$, then the source will always proceed to state $i+1$ since no packet transmissions occur until the guaranteed slot after d frames. After d frames, the source transmits again. This packet marks the beginning of a talkspurt if $S \leq dT$. Hence,

$$\begin{aligned} p_{d0} &= P\{X(t+1) = 0 | X(t) = d\} && (3.16) \\ &= P\{S \leq dT\} && (3.17) \\ &= 1 - e^{-\lambda dT} && (3.18) \end{aligned}$$

It is a simple matter to find the Markov chain for the stationary probabilities π_i. Since $1/\mu$ and $1/\lambda$ are on the order of one one second and T on the order of 10 msec, we make the approximations $e^{-\mu T} \approx 1 - \mu T$ and $e^{-\lambda dT} \approx 1 - \lambda dT$. In this case, the probability that a user has a packet to send that is the result of being in the active state is

$$p = \pi_0 = \frac{\lambda}{\lambda + \mu} \tag{3.19}$$

which is consistent with the analysis of the continuous time two state Markov chain for voice activity. The probability that the source has a packet to send in an arbitrary slot is

$$b = \pi_0 + \pi_d = p + (1-p)\frac{1}{d} \tag{3.20}$$

This result is intuitively satisfying since a user that is inactive with probability $1-p$ will offer a packet once every d frames.

Chapter 4
Voice and Data Integration

4.1 Introduction

Packet data techniques provide an efficient means for sharing communication resources among users with time varying demands. The application of packet techniques have been mainly in the digital data communications where the bursty nature of the traffic is exploited to gain large efficiency advantages in utilization of communication resources. Packet techniques exhibit important benefits for voice (as was discussed in previous chapters) as well as for data.

The integration of voice and data in a common packet switched radio network offers a number of potential benefits, including efficient utilization of the transmission resources, and enhanced services for users requiring access to both voice and data communications.

We now extend the DRMA protocol for the integration of voice and data. In DRMA, the data traffic will be made to use the silence intervals in the voice calls effectively. It should be noted that priority is given to the voice users. Whatever scheme is the most suitable one, it is made sure that only the bandwidth that is going waste, i.e, the bandwidth voice users do not make use of is allocated to the data users. Thus, no matter what the data user access scheme is, or what their performance is, the data users do not affect the voice users. However, the converse does not apply. The data users' performance will depend on the number of voice users in the system. Greater the number of voice users, lesser is the excess bandwidth, and hence fewer data users can be supported for a fixed performance level, or else the same number of data users at a lower performance level.

A number of multiple access schemes could be used to transmit data. For the data traffic, it is important that a proper scheme is chosen; however, this task does not have a precise solution. The type of access scheme chosen depends on the type of data being considered. In this thesis we focus mainly on email messages, i.e., relatively short messages. A brief description of different access schemes and the most suitable one for the type of data messages considered here, i.e., email messages, is given in the following section.

4.2 Data Access Schemes

We shall now look into two types of data access schemes, polling and random access. Two variation to the polling scheme are considered. We will see that the random access may be the better suited scheme for the type of data considered here.

4.2.1 Polling Schemes

DRMA for Data

If we implement DRMA for the data users, each data user is polled once over a fixed polling interval of T frames. If user 1 is polled in frame t_1, then he would be polled next in frame $t_1 + T$. Every user is polled at an interval of time equal to the fixed polling period. See Figure 4.1. It should also be noted that since the data users are guaranteed to be polled once every T frames, there is no need for a separate control channel.

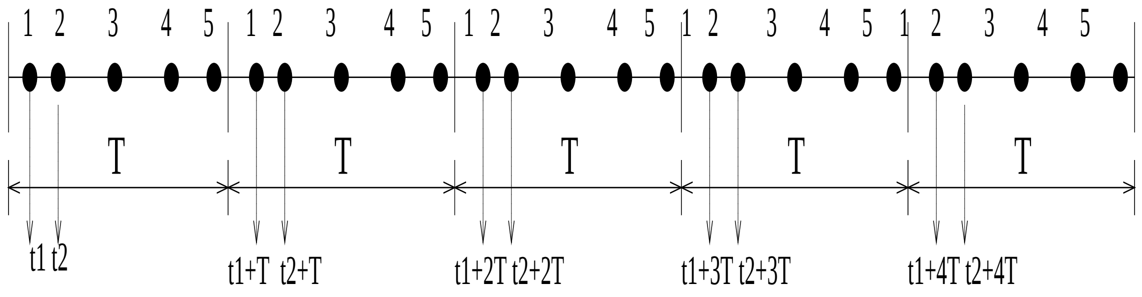

Figure 4.1: DRMA for data users.

There are a few points to be noted. In frame $t_1 + xT$, if all the slots are occupied by either voice or data users or a combination of both, what is the possible course of action? If there are one or more than one data users using slots in frame $t_1 + xT$, one of the data users is preempted and the slot is given to user 1 for polling. However, a voice user should not be disturbed. In the case where all the slots are occupied by voice users, data user 1 would have to wait till the next frame where there is bandwidth available to be polled. This bandwidth could either be an idle slot, or another data user could be preempted and the slot assigned to poll user 1 in that frame.

Message (data) arrival

We assume that the message arrivals from the source is a Poisson process of rate λ. We would like to find the average number of packets that would be available at the data source at the end of the polling period of duration T.

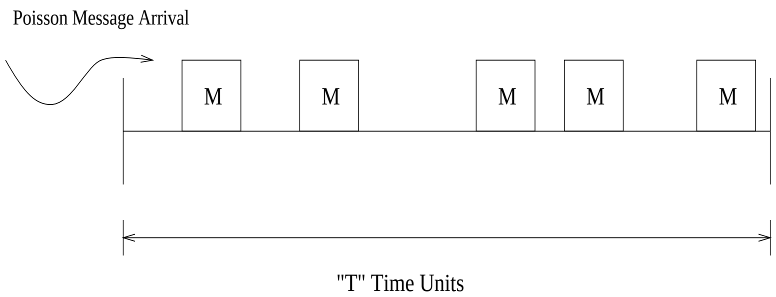

Figure 4.2: Poisson Message Arrival

We will call the collection of backlogged packets per polling interval of a data user, a burst. Let X_i denote the length in packets of the i^{th} message. The number of packets in a burst is,

$$R = X_1 + X_2 + X_3 + \cdots + X_N \tag{4.1}$$

Note that R is the sum of N random variables $X_1, X_2, \ldots, X_N$. The expected number of packets that arrive in the interval is

$$E(R) = E(X)E(N) \tag{4.2}$$

We now characterize random variable X. It has been found through experiment, which will be explained further on in this subsection, that it is reasonable to model X by a geometric random variable. Therefore,

$$P_X(x) = \mu(1-\mu)^{x_0-1} \tag{4.3}$$

$$E(X) = \mu^{-1} \tag{4.4}$$

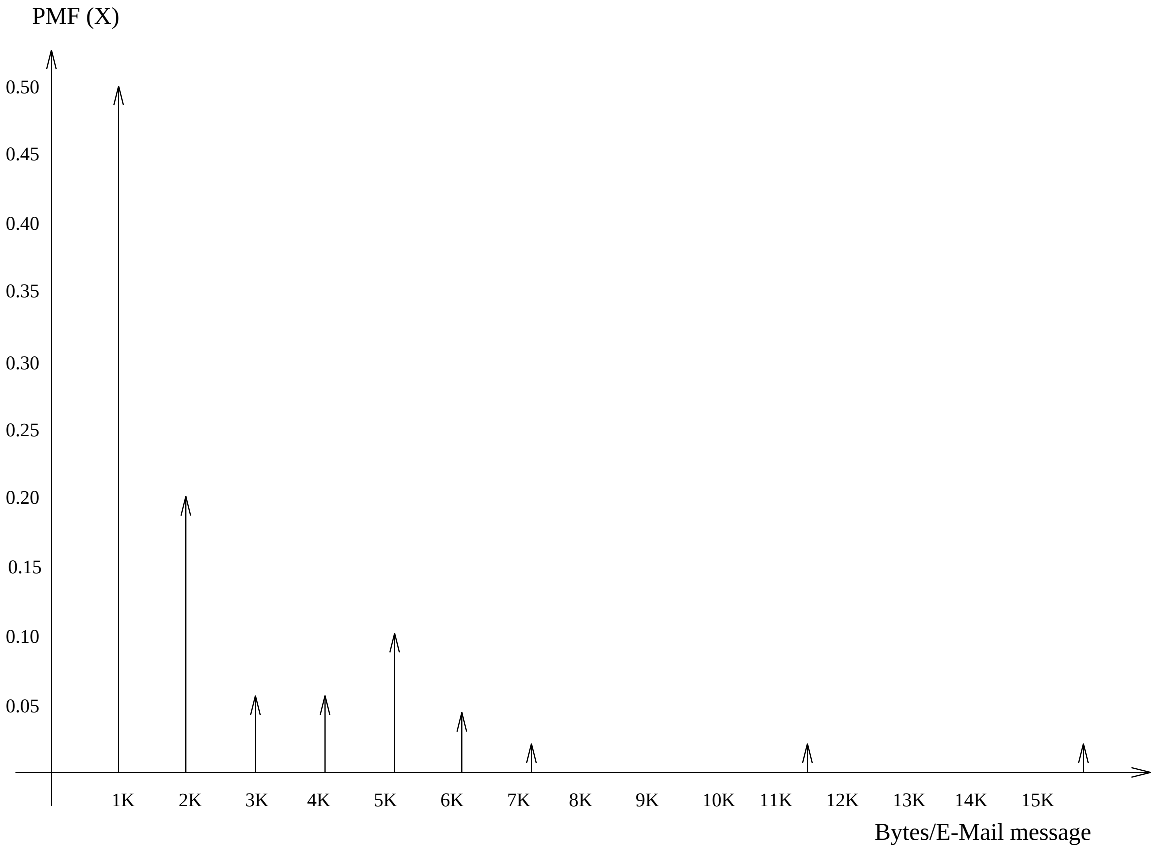

Figure 4.3: Histogram of a sample of E–mail messages.

n_0	$P_N(n_0)$
0	0.594
1	0.309
2	0.080
3	0.013
4	0.0018

Table 4.1: Probability of arrival of n_0 messages in a data user's queue at the end of a polling period.

The number N of message arrivals is Poisson distributed, therefore,

$$P_N(n) = \frac{(\lambda T)^n}{n!}\exp[-\lambda T] \qquad \text{n} = 0,1,... \tag{4.5}$$

$$E(N) = \lambda T \tag{4.6}$$

where T is the duration of the polling interval. Using the geometric model for X, we have,

$$\sigma_X^2 = \frac{(1-\mu)}{\mu^2} \tag{4.7}$$

The variance of random variable R can be found from the formula below [1],

$$\sigma_R^2 = E(N)\sigma_x^2 + [E(X)]^2\sigma_N^2 \tag{4.8}$$

An experiment on a sample of E-mail messages has led us to the conclusion that X could be reasonably approximated by a geometric random variable with a mean of $E(X) = 30$ packets/message, where each packet is 50 bytes long. See Figure 4.3. For example, if we consider that the arrival rate λ is 50 messages per day i.e., 0.0347222 messages/min and the polling interval T is 15 minutes, then, $E[R] \approx 15$ packets/burst and $\sigma_R \approx 26$ packets.

The probability that there are $N = 0, 1, 2, \ldots$ messages in the queue for a single data source at the end of the polling period of duration $T = 15$ minutes can be seen from the Table 4.1. It is seen that, if the data sources constitute only e-mail sources, the amount of data is very low and the probability that a source may have a message to transmit at the beginning of the new polling period is very low.

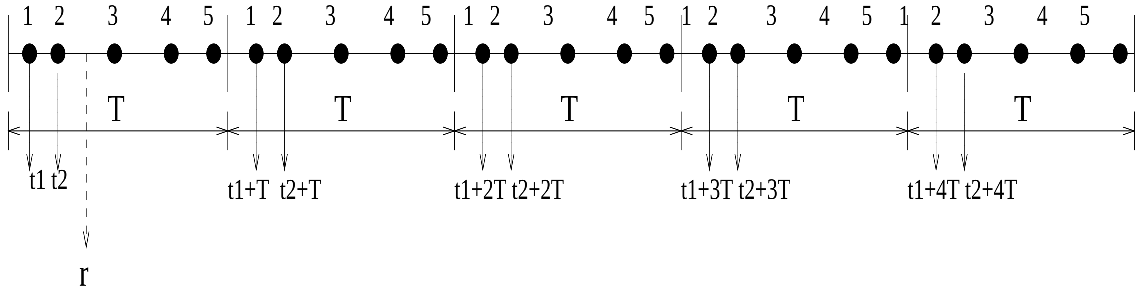

Figure 4.4: An illustration of possible waste in bandwidth due to DRMA for Data.

Polling Data Users on Available Bandwidth – Variable polling interval

In case of the DRMA for data, if users 1 and 2 complete transmitting their data before any other users are polled, say at a time γ as shown in the figure 4.4, then, provided that there was any idle bandwidth available after the voice users, it would all go unused till the next user is polled. This could be avoided if the users are polled sequentially only when there is available bandwidth, i.e., users are not polled at fixed intervals but when there is BW available. The moment an idle slot is available, the next data user in sequence would be asked (by the BS on the downlink) to transmit if he has any messages. This method is commonly known as the round robin method.

In the two polling methods discussed above, since the users are checked sequentially, it could result in long delays when there are large number of data users in the system. It does not make much sense to poll the data users as in DRMA, since, when the data user is polled next time after an interval of T frames, if there is no available idle bandwidth (all slots in the frame occupied by voice users), he does not get to transmit until the next frame in which there is bandwidth available. In both the methods above, one slot is wasted if the data user does not have anything to transmit.

Thus it is seen that both the polling schemes are inefficient if viewed from the delay point of view. The delay is especially very large when there are a large number of data users in the system. It can be seen from the table 4.1 that for low arrival rate (such as for email messages) the probability that a user has a message every polling interval is very low. This will result in a large number of slots being wasted for polling the data users, thus causing a large delay to other data users with information to transmit. If polling scheme **has** to be used, then "polling on available BW" may be better than using "polling at a fixed interval". Another important point is that polling data users requires mobility management, since the mobile users must be tracked. The polling period for the data users is long relative to that for voice users and this may result in the data user crossing over to another service area, thus necessitating tracking of the data user. An alternative to polling schemes is to use the **random access schemes**. Here, only those users who have messages to transmit, contend to gain access to the channel. This could result in a smaller delay to the users than in the polling strategy, where a data user with a message has to await his turn irrespective of whether the users to be polled ahead of him have messages or not.

4.2.2 Random Access Method

In the random access method, once a data user gains access to the slot, he reserves the slot until he finishes transmitting the entire message. Once he finishes transmitting, he relinquishes control over the slot. This is similar to the slotted Aloha protocol with reservation strategy. However, this contention method also results in a delay due to colliding users i.e., users who contend for the same slot at the same time. When two or more users collide, they back off, wait a random amount of time independently (which is a function of the contention probability p) and contend for a slot again. If there are a large number of users there could be a large number of collisions and hence a large delay. Even with only a few data users in the system, under the condition that no packets are to be dropped and with no contention resolution, the system would slowly become unstable at some point of time [4] and the delay will go to infinity. Therefore some kind of contention resolution scheme is required in order to make the system stable under the condition that the load is less than or equal to a certain λ_{max}.

Unlike reservation Aloha, where the slot that is reserved by the data user could be used by the data user in every frame till the end of the talkspurt, in DRMA a data user may have to relinquish control over a slot he reserves at least temporarily, when a voice user requests to transmit in that slot, thus giving priority to the voice user over the data user. Also, in DRMA the data users will not be able to contend in every slot as they may be occupied by a voice user or a data user who reserved the slot prior to the current contention. When the system contains voice users as well as data users whose messages contain more than one packet and the data user reserves the slot till it transmits all its packets, idle slots do not appear continiously as slots are usually occupied by either the voice users or other data users. At the beginning of each frame, the base station informs on the broadcast channel, the status of each slot, so that the data users could know which slots to contend in. The data protocol could also be used by the voice users for the initial access of the system.

Collision Resolution Algorithm

Slotted Aloha requires some care for stabilization and is also limited to throughputs of $1/e$ [4]. The throughput may be significantly higher in the case of reservation Aloha. However, without stabilization, the system backlog will eventually grow to be unacceptable.

We therefore look at a collision resolution technique that maintains stability without any complex backlog estimation procedures. By a collision resolution algorithm, we mean a protocol for the transmission and retransmission of messages by all individual nodes with the property that after each collision all nodes involved in the collision eventually gain access to the channel and become aware that these nodes have gained access to a channel. A collision is resolved at the point where all the nodes become aware of the fact that all the colliding nodes have gained access to the channel.

When a collision occurs, all the new arrivals could be inhibited from transmission until the collision is resolved. The new messages are considered as backlogged and they get to transmit only after the collision is resolved. There are many ways in which a node involved in a collision could choose whether or not to transmit in the successive idle slots. Each node could simply flip an unbiased coin for each choice or each node could use the arrival time of its collided message. The set of colliding nodes are split into subsets, one of which transmits in the next available slot and thus reserves the slot till the end of the message. The nodes in the other subset transmit after the collision (if any) in the previous subset is resolved. Algorithms of this type are known as the splitting algorithms [4]. The particular algorithm we look into is known as the Capetanakis Collision Resolution Algorithm, also commonly known as the Tree Algorithm.

The following assumptions are made:

• The channel is a time-slotted collision-type channel, but is otherwise noiseless.

• The feedback channel from the receiver or the base station is a noiseless broadcast channel that informs the nodes immediately at the end of the slot whether 1) the slot was empty i.e., idle 2) a packet was transmitted i.e., success 3) a collision of two or more nodes i.e., collision. This type of feedback is also known as $(0, 1, e)$ feedback.

• Propagation delays are negligible so that the feedback information is immediate and this information could be used to determine as to who should transmit in the following slot.

Capetanakis Collision Resolution Algorithm or the Tree Algorithm

In this section we shall first explain the CCRA in a general fashion followed by an example which should clarify the algorithm and illustrate its main features. The period between the start of the collision resolution and the end of it is commonly known as the collision resolution interval (CRI) or the collision resolution period (CRP).

Suppose a collision occurs in some slot n, all those nodes not involved in the collision go into a waiting mode. Those nodes involved in the collision split into two subsets (by flipping a binary coin). The first subset transmits in slot $n+1$, and if that slot is idle or successful, the second subset transmits in slot $n+2$. Alternatively, if another collision occurs in slot $n+1$, the first subset splits again and the second subset waits until the collision resolution of the first subset.

The Figure 4.5 represents a binary tree structure. In the Figure, "C" represents a collision, "S" represents a success and the "I" represents an Idle. The "A" represents an initial set consisting of colliding nodes. All the colliding nodes split into the two subsets, the upper and the lower. All the nodes in the upper subset then get to transmit in the next slot. All the nodes in the lower subset wait until the collisions (if any) in the upper subset are resolved. Each vertex in the tree corresponds to a subset of nodes involved in the collision. Vertices whose subsets contain two or more nodes contain

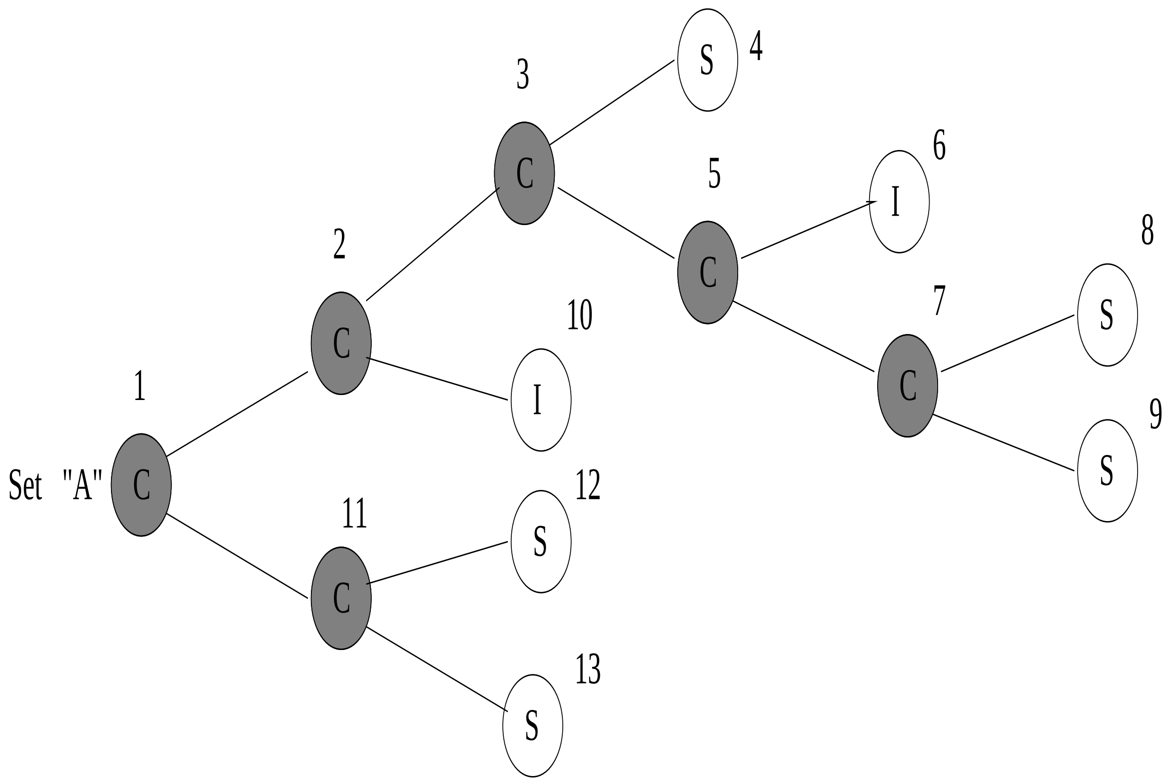

Figure 4.5: CCRA or the Tree Algorithm.

two branches corresponding to the upper and lower subsets. Vertices corresponding to subsets consisting of 0 or 1 node are leaf vertices of the tree. Given the immediate nature of the feedback, each node can construct the tree as shown in 4.5 as the 0, 1, e feedback occurs, thus keeping track of it's own subset in the tree so that it can determine for itself, when to start transmitting its own backlogged message.

The CCRA can be defined as follows [15],

> After a collision, all the nodes involved in the collision flip a binary coin. Those flipping 0 transmit in the very next slot, those flipping 1 retransmit in the next slot after the collision (if any) among those flipping 0 has been resolved. No new arrivals may be transmitted until after the initial collision has been resolved.

We now explain the CCRA with the help of an example. Let us say that the initial collision is among 4 nodes, A, B, C and D. These 4 nodes flip their binary coins at the end of slot 1. Let us suppose that A and C flip 0 while B and D flip 1. Now A and C

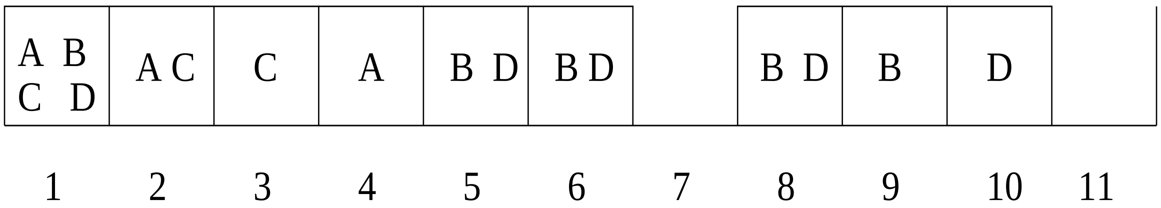

Figure 4.6: Example of a Collision Resolution Interval for the tree algorithm.

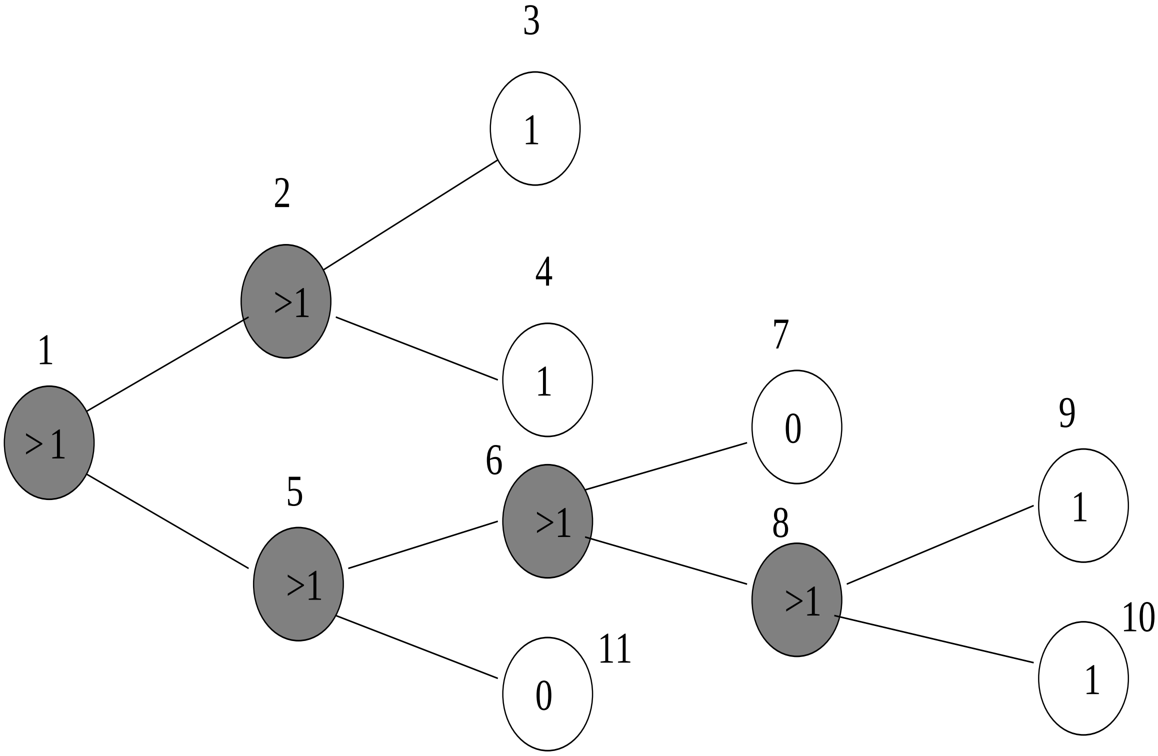

Figure 4.7: Tree diagram for the CRI of the previous figure.

try to transmit in slot 2 and it results in a collision. A and C flip coins again. Let us say A flips 1 while C flips 0, then C alone gets access to the channel in slot 3 while A gets access to the channel in the next available idle slot.

The numbers inside the nodes in the Figure 4.7 represent the number of users in that subset. The number on top of the node represents the slot number. The Figure 4.6 shows the CRI for CCRA. In Figure 4.6, the collision in slot 2 has finally been resolved in slot 4. Now, the nodes B and D which have been waiting for the resolution of the collision in the upper subset transmit in slot 5. At the end of slot 5, B and D flip coins again. Say both the nodes flip 0. Then B and D end up in the upper subset and the lower subset in empty. In slot 6 again, there is a collision; B and D flip coins at the end of slot 6. Say they both flip 1, resulting in a collision in slot 8. The next flip results in say, B flipping 0 and D flipping 1. Then, B gains access to the channel in the 9th idle slot and D in the 10th idle slot. Thus we would expect that the collision resolution would end at this point. However, it does not, because, the receiver or the base station does not know whether or not there are users in the empty lower subset created earlier. Thus, slot 11 is used by the empty lower subset. Since this subset is empty (as can be seen from the figure 4.7) the slot goes empty. All the nodes (not just the ones involved in the collision) can grow the tree in figure 4.7 from the feedback information, and thus all nodes now simultaneously learn that all the nodes in the original collision have gained access to the channel.

Implementation of CCRA

When a node flips a 1 following a collision in which it is involved, it sets the counter to 1, then increments it by one for every subsequent collision slot and decrements it by one for every subsequent collision free slot. When the counter reaches 0, the transmitter retransmits it in the next slot [15].

All nodes must also know when the original collision has been resolved, so that all the new arrivals that were backlogged during the contention resolution, could now be transmitted. This could be done with the help of a second counter which is set to 1 just before the first slot, then incremented by one for each subsequent collision slot and decremented by one for each subsequent collision free slot. When the counter reaches 0, it implies that the collision has been resolved. The newly arrived backlogged messages can now transmit in the next slot. The period between the start of the collision resolution and the end of it is commonly known as the collision resolution interval (CRI) or the collision resolution period (CRP).

It can be observed that in the implementation of the CCRA, the feedback information need only determine whether the corresponding slot has a collision or is collision free. Therefore the feedback need not differentiate between a successful and an idle slot.

4.3 Simulation of the Data Protocol

A brief explanation of the data protocol is first made which is followed by a summary of the results obtained through the simulation. We shall then look into the analysis of the throughput for this data protocol in the following section.

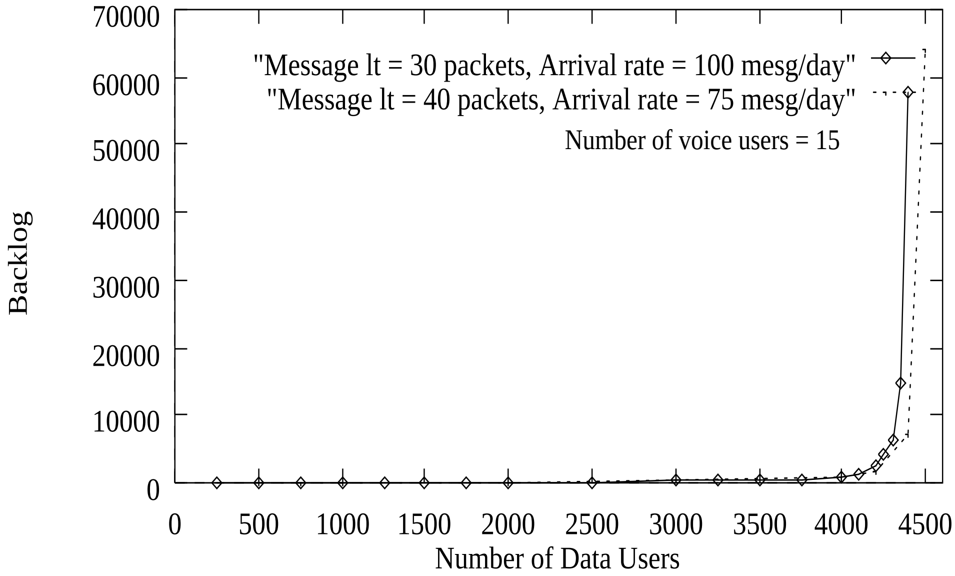

Figure 4.8: Backlog in terms of the number of packets as a functio n of the number of Data Users

The DRMA protocol is designed such that voice users are given priority over the data users. The data users only make use of the excess bandwidth that is available after the voice users. The data users contend for channel access when they have a new arrival. They can only contend for the channel when the slot is not occupied by an active or an inactive voice user (guaranteed slot) or an active data user who has already reserved the channel earlier. If there is only one data user contending in an idle slot, the data user gains access to the channel and reserves that slot for the entire duration of the message. If there are two or more users contending to gain access to a channel in the same slot, it would result in a collision and collision resolution has to take place before any other new arrivals are allowed to transmit. All the messages that arrive during the CRI are considered as backlogged as has been explained in the previous section i.e., we implement the CCRA or the Tree Algorithm in order to resolve the contention.

It can be seen from figure 4.8 that the data delay starts to blow up as the number of users cross 4000 with each user having an arrival rate of 100 messages/day with each message having an expected length of 30 packets. This translates to a load of about 1.52 packets per frame. The exact upper bound on the throughput is calculated in the following section. However, 1.52 packets per frame appears to be some what conservative in comparision to the throughput obtained through the simulation. In simulation the system appeared to be stable for 4250 users which translates to about 1.62 packets per frame. The system became clearly unstable somewhere between 4250 and 4500 users for the above mentioned message arrival rate. It can also be seen that as the message length (in terms of expected number of packets per message) increases, the protocol performs better. That is fairly obvious as more packets are successfully transmitted per CRI as the lengths of the messages become larger. We will see from the analysis in the following section, the relation between the throughput λ and the expected message length.

4.4 Analysis of Data Throughput

The data protocol is a random access protocol that makes use of the CCRA algorithm in order to resolve the contention. The data user also reserves the slot it acquires during the CRI till all the packets in its message are transmitted. Therefore, both these factors i.e., the expected CRI length as well as the expected length of the message influence the stable load that could be handled by the protocol. We now analyze the maximum stable load of the data protocol following the method of Massey [15].

We would first like to find the mean CRI length. First and the most important is the conditional mean CRI length, A_n. The CRI length is conditioned on the number of users, n, who contend in the initial slot of the CRI. Let us represent this quantity by A_n, i.e.,

$$A_n = E[C|X = n] \tag{4.9}$$

where C denotes the length in slots for the CRI to be resolved and X denotes the number of packets contending in the first slot of the CRI.

We would now like to determine the dependence of the mean CRI length on n. Let us first assume that n is a very large number. If a coin is flipped in order to determine which of the two subsets (upper or lower) each user goes into, the probability that i of the colliding users would be in one or the other subsets is [15],

$$p_n(i) \quad = \quad \binom{n}{i}\left(\frac{1}{2}\right)^i\left(\frac{1}{2}\right)^{n-i} = \binom{n}{i}2^{-n} \tag{4.10}$$

The expected CRI length is equal to $E[1 + A_i + A_{n-i}]$, ie.,

$$A_n = 1 + \sum_{i=0}^{n}(A_i + A_{n-i})p_n(i) \tag{4.11}$$

but, $p_n(i) = p_n(n-i) = \binom{n}{i}2^{-n}$, therefore,

$$A_n \quad = \quad 1 + 2\sum_{i=0}^{n} A_i p_n(i) \tag{4.12}$$

$$\quad = \quad 1 + 2[\sum_{i=0}^{n-1} A_i p_n(i)] + 2A_n p_n(i) \tag{4.13}$$

Thus,

$$A_n = \frac{1 + 2[\sum_{i=0}^{n-1} A_i p_n(i)]}{1 - 2^{1-n}} \tag{4.14}$$

Solving the above equation for various values of n with $A_0 = A_1 = 1$ as the initial values, we obtain the values shown in table 4.2. From this table it can be seen that, $A_n - A_{n-1} \approx 2.9$ for $n \geq 4$. It can also be seen that $A_n = 2.9n - 1$ is quite accurate for $n \geq 3$. A tighter bound as found by Massey [15] is $A_n \approx 2.89n$. In our protocol, the mean CRI length is greater than $2.89n$ due to the presence of the voice users in the system. We shall however, take into account the effect of voice users later on in the discussion. For now, let us assume that there are only data users in the system and proceed with the analysis.

Let $E[L]$ be the expected message length and λ be the message arrival rate. If n messages were to contend in the initial slot of the CRI, the expected number of slots required for the contention to be resolved (i.e., the expected CRI) and the messages to be transmitted is,

$$E[S] = 2.89n + [E[L] - 1]n \tag{4.15}$$

The expected number of packets sent in E[S] slots is,

$$E[R] = E[L]n \tag{4.16}$$

n	A_n	$2.9n - 1$
0	1	-
1	1	-
2	5	-
3	7.667	7.7
4	10.524	10.6
5	13.419	13.5
6	16.313	16.4

Table 4.2: Expected CRI length for the CCRA given n packets in the first slot.

Therefore, the effective throughput in packets/slot is,

$$\lambda_{max} = \frac{E[R]}{E[S]} \tag{4.17}$$

$$= \frac{E[L]n}{2.89n + [E[L] - 1]n} \tag{4.18}$$

For the system to be stable, the maximum arrival in packets/slot must obey,

$$\lambda_{max} \leq \frac{E[L]}{2.89 + E[L] - 1} \tag{4.19}$$

This is the throughput bound for the case in which there are absolutely no voice users in the system. However, in a voice and data integrated system such as DRMA, not all slots are available for data. Voice is given priority over data. The excess bandwidth left over after the voice users is given to the data users, explicitly giving priority to the voice users. We would like to calculate the excess bandwidth under the constraint of a fixed number of voice users in the system.

The probability that a voice user has a packet to send in a frame is, from equation 3.1,

$$b = p + \frac{(1-p)}{d} \tag{4.20}$$

where p is the voice activity factor, d in the second term is the maximum packet delay constraint as measured in frames. In particular $\frac{1}{d}$ is the packet transmission rate, in slots per frame, of an inactive user.

With K voice users in the system, the expected number of voice users with packets to send in a frame is,

$$E[V] = Kb \tag{4.21}$$

and,

$$P\{V = i\} = \binom{K}{i} b^i (1-b)^{K-i} \qquad \text{i} = 0,\ 1, \cdots, \text{K} \tag{4.22}$$

The expected number of slots idle in a frame i.e., the excess bandwidth left behind the voice users could be calculated as $N - E[V]$, where N is the number of slots per frame. Thus the data protocol integrated with the voice users is stable under the condition that

$$\lambda_{max} \leq \frac{E[L]}{2.89 + E[L] - 1}(N - E[V]) \text{ packets/frame} \tag{4.23}$$

The same analysis could be done at the superframe level too. The results obtained from this work out to be the same as those obtained from the above equation. We shall see that soon. Now the excess bandwidth could be calculated as the expected number of idle slots per superframe. The expected number of idle slots per superframe could be calculated as $E[(Nd - dS_A - (K - S_A))^+]$, where d is the number of frames per superframe and $E[S_A]$ is the expected number of active users in any arbitrary superframe.

$$E[S_A] = KP_{on} \tag{4.24}$$

where P_{on} is the probability that a user is in an active state in any arbitrary superframe. We can now calculate the maximum throughput of the data users as a function of the message length and the number of voice users as shown below.

$$\lambda_{max} \leq \frac{E[L]}{2.89 + E[L] - 1} E[(Nd - S_A(d-1) - K)^+] \text{ packets/superframe} \tag{4.25}$$

$$= \frac{\frac{E[L]}{2.89+E[L]-1} E[(Nd - S_A(d-1) - K)^+]}{d} \text{ packets/frame} \tag{4.26}$$

It can be seen that the upper bound on the maximum throughput increases with an increase in the size of the expected message length. This is obvious as the fraction of packets transmitted per slot for a CRI of length $2.89n$ is greater as $E[L]$ increases.

The graph in Figure 4.9 shows the maximum throughput for the data users as a function of the number of voice users. The two graphs in the figure represent the maximum throughput per frame for two different values of the message lengths, 30 packets per message and 100 packets per message. It can be seen that the protocol has a higher upper bound on the maximum data throughput when the expected number of packets per message is higher. It can also be seen from figure 4.10 that the throughput

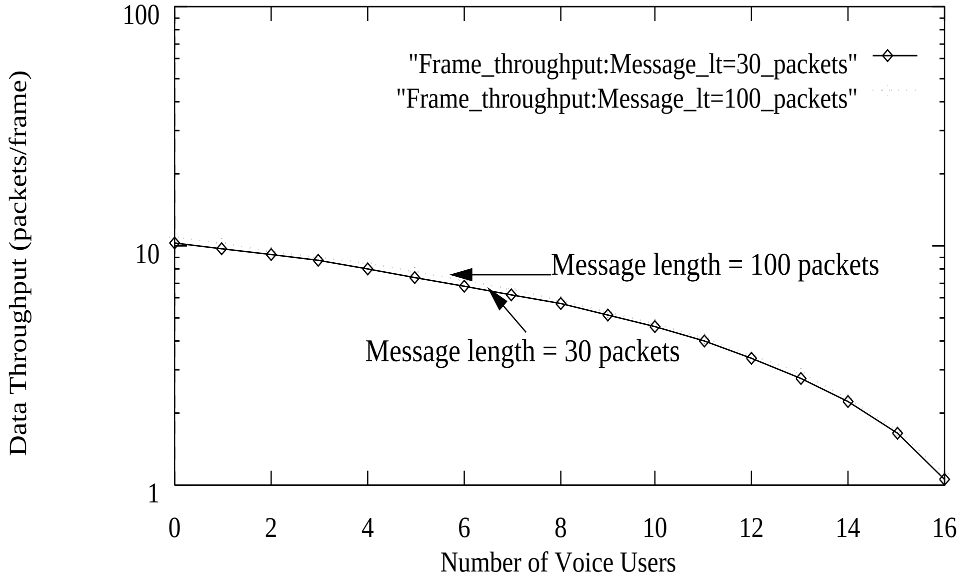

Figure 4.9: Maximum throughput per frame as a function of the number of voice users for two different message lengths.

results obtained from analysis through frame and superframe are almost the same. The graphs obtained in Figure 4.10 are for an expected message length of 30 packets. It can be seen that the data throughput value obtained through analysis when there are 15 voice users in the system agrees with the results obtained through simulation. This can also be seen from Figure 4.11.

Figure 4.11 shows how the throughput changes as a function of the message length for a fixed number of voice users. In this particular figure the number of voice users is fixed at 15. It can be seen that the throughput increases a lot between message lengths of 1 and 15 packets after which it almost saturates. This is pretty obvious as the increases in the ratio of the message length to the CRI length+message length (as can be seen from equations for λ_{max}) becomes very small as $E[L]$ increases to a large value.

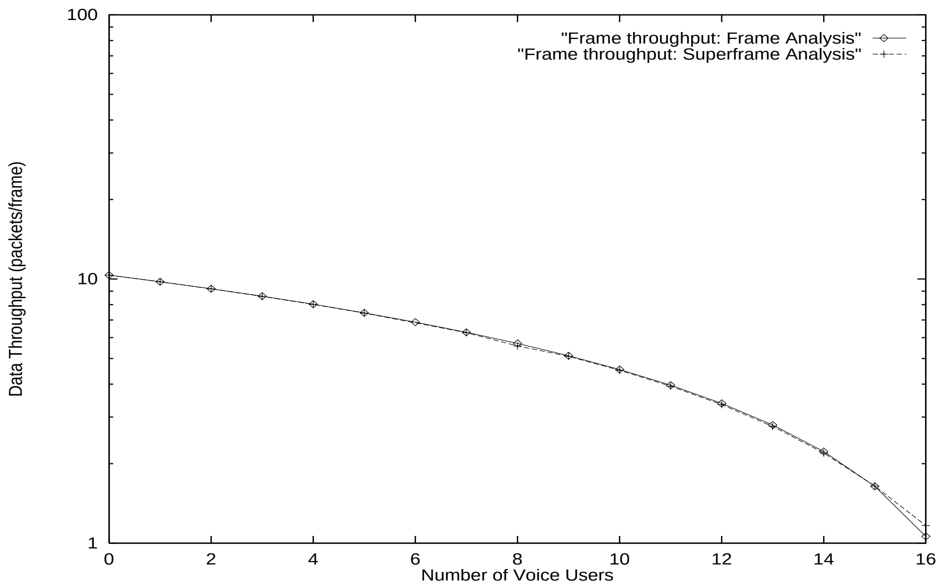

Figure 4.10: Results of throughput per frame obtained through frame level and superframe level analysis.

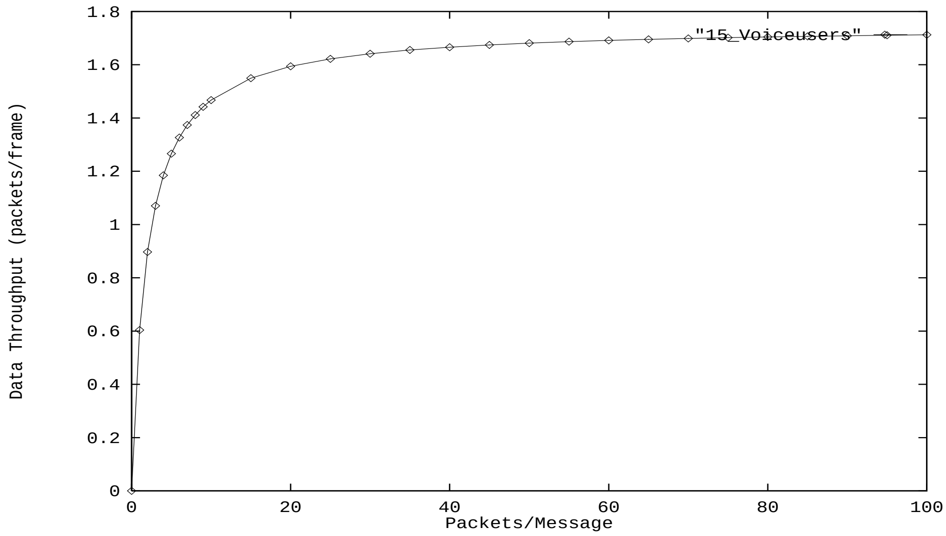

Figure 4.11: Throughput per frame as a function of the expected message length, with number of voice users fixed at 15 i.e, K = 15.

Chapter 5

Conclusion

Work in this thesis has examined a media access control protocol known as Dual Rate Multiple Access in order to improve the spectral efficiency of a cellular/PCS system for voice and data communications for a given physical layer technology relative to the TDMA systems. Chapters 1, 2 and 3 dealt with introduction, description, simulation and analysis of DRMA for voice users only. In chapter 4 we looked into integration of data and voice for the DRMA protocol.

In the first chapter, we have dealt with briefly, the various methods by which we could improve the bandwidth efficiency for a particular system. Improvements in the physical layer technology could lead to very efficient spectral utilization. In this report however, we look into what improvements in spectral efficiency could be afforded at a layer above the physical layer i.e, the MAC layer. Prior work has been done on systems that utilize statistical multiplexing in order to improve the bandwidth utilization. Some of the very significant ones are reservation ALOHA and Packet Reservation Multiple Access (PRMA). Both the above mentioned protocols use contention mechanism as the media access technique. PRMA has a 54.5% increase in spectral utilization as compared with the traditional TDMA techniques. However, PRMA has certain disadvantages such as the *dead line* problem and the need for a control channel on a separate band. DRMA counters these problems with the use of polling as the media access technique.

The second chapter deals with the description and simulation model of the DRMA protocol. We explain in detail, the frame structure and the manner in which different users are assigned to different slots as the state of the system changes dynamically. During the course of the simulation, a significant effect known as the *shift in the guaranteed slot sequence* has been observed. This effect has been observed through simulation of

DRMA for voice. This has also been explained in the second chapter through the help of an example. This ability to advance the guaranteed slot results in a lower level of packet dropping that is observed in the frame level analysis, where it is assumed that the position of the guaranteed slot remains fixed. The second chapter also covers the effect of buffering packets when there is no high rate channel available for the transfer of voice packets. It has been found through simulation, that, buffering also leads to an improvement in the packet dropping rate, which is quite significant when the number of simultaneous users in the system are small but gets trivial as the number increases.

In chapter 3, performance of the DRMA protocol has been studied through analysis. This is then followed by a comparisison of the analysis with the simulation model. Analysis of the protocol is done at two levels. In the first case, analysis is done at a frame level and in the second method, analysis is done at the superframe level. In the latter case it is assumed that the packets are sent within the superframe and assumes that a low rate user is alloted any one slot in the superframe so that the guarenteed slot position may not be fixed, as is the case in the first analysis method. It is seen that the superframe analysis method is closer to that of the simulation model and is thus more realistic. Before concluding this chapter, we gave a model of the DRMA source activity. It has been found through analysis as well as simulation that, DRMA can support 15 users as compared to the 11 that a TDMA scheme could support when the number of slots is 11 per frame, thus, affording affording a 36% gain in spectral efficiency when compared with TDMA schemes. DRMA not only provides a gain in spectral efficiency, as is the case with PRMA, but also provides a low bit rate representation of the background noise/dedicated control channel, unlike PRMA. We therefore conclude that, inspite of a slight decrease in spectral utilization in DRMA in terms of the number of simultaneous calls, as compared with PRMA, DRMA may be the more efficient method as it eliminates the *deadline* and provides two rates of service, in doing which it provides an embedded full time control channel within the same band.

In the fourth chapter, we looked into the requirements, methods, realizable models for voice and data integration in the DRMA protocol. A number of methods for media access, such as polling, round robin, contention etc., have been considered. It was finally concluded that random access may be the most suitable method of media access for the data users. It is to be noted that, while the data users method of media access is contention based, the voice users still use polling as their way of accessing the media. Also the data users get to use only the residual bandwidth in order to communicate, thus giving explicit priority to the voice users over the data users. Random access of the channel by the data users results in collisions between the users and hence requires collision resolution techniques in order to make the protocol efficient. Here, we looked into Capetanakis Collision Resolution alogorith (CCRA) or what is commonly known as the Tree algorithm. A brief description of the implementation of CCRA and the simulation of the protocol is then followed by an analysis of the data throughput for the integrated voice and data DRMA protocol.

5.1 Future Work

The parameters that have been used in the system, such as the frame length and the number of slots, are in reality, parameters that have been optimized for the PRMA system. The exact parameters to be used in order to optimize the DRMA sytem is one area open for research. Also, the quality of voice was found to be good at 1% packet dropping through subjective tests or mean opinion score (MOS) tests particularly for the PRMA system. In DRMA, since background noise during silence periods is also included, the voice quality may be acceptable even at packet dropping probabilities higher than 1%. This is another topic that needs to be examined.

In the integration of voice and data, we have considered data messages to be short email messages. Under this assumption we have concluded that random access may be the best suited method for media access by the data users. We then performed the analysis of the data throughput for the DRMA protocol. However, if the data messages are much longer and their arrival rates are unlike that of the Email messages, then random access may not turn out to be best. Though we could never say that any particular method of access is the best for data (considering that each method has its own advantages and disavantages based on the type of data), further work needs to be done to identify appropriate access methods for differing types of data messages.

References

[1] Drake, A.W. *Fundamentals of Applied Probability Theory.* McGraw-Hill Book Company, 1967.

[2] Tanenbaum, A.S. *Computer Networks.* Prentice-Hall, 1992.

[3] Bianchi, G., Borgonova, F., Fratta, L., Musumeci, L., Zorzi, M. "C-PRMA: The Centralized Packet Reservation Multiple Access for local wireless communications". *Proc. IEEE GLOBECOM '94.*

[4] Bertsekas, D.P. and Gallager, R.G. *Data Networks.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1993.

[5] Bratley, P., Fox, B.L., Schrage, L.E. *A Guide to Simulation* Springer - Verlag, 1993.

[6] Chien, E.S.K., Goodman, D.J., Russel, J.E., Sr. "Cellular Access Digital Network (CADN): Wireless Access to Networks of the Future". *IEEE Communications Magazine*, Vol.25, No.6, June 1987.

[7] Calhoun, G. *Digital Cellular Radio.* Artech House, Inc., 1988.

[8] Goodman, D.J., Nanda, S. & Timor, U. "Theory of Packet Reservation Multiple Access". *Technical Report*, WINLAB-TR-2, September 1989. Rutgers University, New Jersey.

[9] Goodman, D.J. "Trends in Cellular and Cordless Communications". *IEEE Communications Magazines*, Vol.29, No.6, pp. 31-40, June 1991.

[10] Goodman, D.J., Pollini, G., Meier-Hellstern, K.S. "Network Control for Wireless Communications". *IEEE Communications Magazine*, Vol. 30, No. 12, pp. 116-124, Dec 1992.

[11] Goodman, D.J., Valenzuela, R.A., Gayliard, K.T., and Ramamurthi, B. "Packet Reservation Multiple Access for Local Wireless Communications". *IEEE Transactions on Communications*, Vol. 37, No. 8, August 1989.

[12] Goodman, D.J., Wei, S.X. "Efficiency of Packet Reservation Multiple Access". *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, Feb 1991.

[13] Gurber, K., and Strawczynski, L. "Subjective effects of Variable delay and Speech clipping in dynamically managed voice systems". *IEEE Transactions in Communications*, Vol. COM-33, pp. 801-808, Aug 1985.

[14] Macario, R.C.V. *Personal and Mobile Radio Systems.* Peter Peregrinus Ltd., London, United Kingdom, 1991.

[15] Massey, J.L. *Collision Resolution Algorithms and Random Access Communications (Report UCLA-ENG-8016).* Los Angeles: University of California, 1980.

[16] Parsons, J.D., and Gardiner, J.G. *Mobile Communication Systems.* Blackie, Halsted Press, division of John-Wiley and Sons, Inc., 1991, NY.

[17] Lee, W.C.Y. *Mobile Cellular Telecommunications Systems.* McGraw-Hill Book Company, 1989.

[18] Wong, W.C, Goodman, D.J. "A Packet Reservation Multiple Access for Integrated Speech and Data Transmission". *Technical Report*, WINLAB-TR-20, September 1991. Rutgers University, New Jersey.

[19] Wong, W.C. "Packet Reservation Multiple Access in a Metropolitan Microcellular Radio Environment". *IEEE Journal on Selected areas in Communications*, Vol. 11, No. 6, Aug 1993.

[20] Stallings, W. *Data and Computer Communications.* Macmillan Publishing company, 1994.

[21] Ishizuka, T; Noguchi, A; and Hanada, E. "Evaluation of speech quality for a packetized speech communication system using multipulse coding". *NEC Research and Development.*, Vol.32, No.2, Apr 1991.

[22] Mouly, M; Pautet, M.B. "The GSM System for Mobile Communications" *International Standard Book Number: 2-9507190-0-7*