MASTER'S THESIS

Reservation Based MAC Protocol for a Wireless ATM LAN

by G. Singh
Advisor: J.S. Baras

CSHCN M.S. 97-1
(ISR M.S. 97-3)

CENTER FOR SATELLITE AND HYBRID COMMUNICATION NETWORKS

The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site http://www.isr.umd.edu/CSHCN/

Abstract

Title of Thesis: Reservation based MAC protocol for a Wireless ATM LAN

Degree candidate: Gagan Singh

Degree and year: Master of Science, 1997

Thesis directed by: Professor John S. Baras
Department of Electrical Engineering

A major issue in the Wireless ATM system design is the selection of a suitable channel sharing/media access control (MAC) technique. The design challenge is to identify a wireless "multimedia capable" MAC approach that provides a sufficient degree of transparency for many ATM applications. In this thesis, a reservation based MAC protocol which satisfies the service requirements of different traffic classes, has been proposed for a centralized network architecture. Each cell in the system consists of a base station (BS) serving many wireless terminals (WTs). The basic idea behind the proposed protocol is a sequence of Control-Data Frames (CDFs); each of which consists of a control phase followed by a data phase. WTs transmitting in the data phase of a CDF also send detailed signaling information to the BS using a new scheme called the "superslot" scheme. The signaling information indicates the number of packets at each WT

and their "priorities". The availability of this information enables the BS to allocate bandwidth such that statistical multiplexing is possible and at the same time ensuring that QoS guarantees are met. An ABR based congestion control scheme that leads to a dramatic improvement in the cell loss performance has also been incorporated in the protocol. Results obtained from a discrete event simulation of the system show satisfactory performance of the protocol.

Reservation based MAC protocol for a Wireless ATM LAN

by

Gagan Singh

Thesis submitted to the Faculty of the Graduate School of the University of Maryland at College Park in partial fulfillment of the requirements for the degree of
Master of Science
1997

Advisory Committee:

Professor John S. Baras, Chairman/Advisor
Dr. Leandros Tassiulas
Dr. M. Scott Corson

Acknowledgments

I wish to express my thanks to some of the many individuals who made this work possible. Foremost among these is my advisor, Professor John Baras, who gave me the opportunity to work on a very interesting and practical problem as part of the research for my thesis.

I am also grateful to Dr. Leandros Tassiulas and Dr. M. Scott Corson for kindly consenting to join the defense committee and review this thesis.

I take this opportunity to thank Samir Hulayalkar with whom I worked closely on this project last summer. I would also like to thank all my friends and everybody in SEIL especially Archan Misra with whom I had several discussions during the course of this work. Finally I wish to thank my family for their love and constant support and encouragement.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Wireless Communication Networks, e.g. GSM, have grown very rapidly over the past few years. While GSM targeted only voice communications, the future wireless communication networks are expected to carry voice, data and video as the wireless traffic is expected to be multimedia in nature. Asynchronous Transfer Mode (ATM) has emerged as the technology of choice for carrying multimedia traffic over B-ISDN networks for wireline network topologies. It is then natural to assume a combination of wireless and ATM-based service at the consumer end of the network[1]. Such a combination will open a big multimedia home or business communication network market.

Existing efforts of building a wireless local area network (LAN) are focussed around the emerging standard of IEEE 802.11 in U.S. and HIPERLAN in Europe. However their development does not take into consideration ATM based service requirements of QoS guarantees for both real time and data traffic. These multimedia applications also require wireless systems like Personal Communication Networks (PCNs) to support a wide range of transmission speeds, flexible connections including asymmetric transmission speed at the forward and the return links, and quality of service control, which conventional Ethernet cannot support. ATM is the only transport technology available that can support mixed multimedia traffic with a single User to Network Interface (UNI) for audio,video and data services on the same medium.

Asynchronous Transfer Mode (ATM) has emerged as the technology of choice for wide area network applications. The main reason for this is because ATM is characterized by its flexibility to accommodate any type of service as well as the high speed transport capability. In fact, ATM technology has been developed not for specific applications but for various types of applications. ATM is capable of carrying any kind of traffic, from circuit switched voice to bursty data traffic, at any speed by changing the transmission rate of 53-byte long ATM cells. ATM is thus becoming very promising and the logical choice for networking multimedia terminals for LANs and WANs alike. Typical applications of multimedia services are listed in Figure 1.1. They include voice, data, still picture and motion video. To support these applications, it is necessary to cover the wide range of service bit rates, from several tens of kilobytes per second to 10 Mbs/second at the highest and to support various service types like CBR (constant bit rate), VBR (variable bit rate) and ABR (available bit rate), and UBR (unspecified bit rate). In addition, the required delay and error rate depend on the application. There is evidently a need to combine the two most promising technologies namely the wireless and the ATM technologies. Wireless ATM is a combination of the two and promises to be a very exciting technology. One of the important issues in a wireless system design is the choice of suitable and efficient medium access

Application	Service Type	QoS	Bit-rate Range
Voice Telephony	CO/CBR	Call blocking permitted Low-med cell loss ok Isochronous	2.4 - 32 Kbps
Digital Audio	CO/CBR	Call blocking permitted Low cell loss reqd Low delay jitter	128 - 512 Kbps
Digital Video Teleconference	CO/CBR or CO/VBR	Call blocking permitted Low-med cell loss ok Low delay jitter	64 - 384 Kbps 1 - 6 Mbps (TV/VCR quality)
Digital HDTV	CO/CBR	Call blocking permitted Low-med cell loss ok Low delay jitter	15 - 20 Mbps
Computer Data	CL Best effort	No call blocking Low cell loss reqd Med delay - jitter ok	0.1 - 1 Mbps
E-mail	CL Best effort	Low transfer rate No call blocking Low cell loss ok High delay ok	9.6 - 128 Kbps
High speed data	CL Burst mode packet data	High transfer rate Very low cell loss reqd Med delay - jitter ok	1 - 10 Mbps

Table 1.1: Typical Application Requirements

strategies, as bandwidth is a limited and a very valuable resource in the wireless medium. Here we explore the design of a "multimedia capable" MAC protocol for a wireless ATM system.

1.2 Contribution of the thesis

In the work here, we present a reservation based MAC protocol for a home user/office worker oriented wireless ATM network, which is capable of supporting various traffic services and also guaranteeing QoS to services which require them.

Most of the existing MAC protocols for wireless LAN systems don't take into account detailed state information namely the time to expiry, buffer overflow probabilities etc, of the packets queued at each mobile in the system. The transmission of such information by the mobiles to the base station dynamically, typically entails a large overhead. However, in our view, the availability of such information is essential for the Base Station/Central Controller to be able to allot bandwidth to each user in an efficient manner and meet the QoS guarantees. The slot allocation scheme in the protocol presented here takes into account the dynamic state information of the packets queued at each mobile. A new signaling scheme called the ATM "superslot" scheme is proposed here, which conveys the dynamic state information of a large number of packets queued at each mobile and still requires very little overhead. The performance of this scheme is compared with other existing schemes using MATLAB simulations. A congestion control scheme based on the ABR service based explicit rate scheme has also been incorporated in the MAC protocol, which leads to a dramatic improvement in the cell loss ratios at high loads. A discrete event simulation of the system has been done using a network simulator BONeS, to evaluate the performance of the protocol. We have also been able to extract certain design parameters of the protocol using this simulation. The protocol has been designed under the assumption that the user in the above-mentioned scenarios requires limited mobility, but good quality of service.

1.3 Organization of the report

Chapter 2 gives an overview of ATM which has emerged as a technology of choice because of its ability to support various classes of traffic. A brief description is given of the various traffic classes that are supported in an ATM network. ABR flow control is also discussed briefly.

Chapter 3 introduces the concept of Wireless ATM, and describes where the usage of such a system is first foreseen. System design criteria and system architecture for a WATM network are also discussed.

Chapter 4 gives an introduction to wide variety of existing medium access strategies and why most of them are unsuitable for a WATM system. A reservation based MAC protocol capable of supporting QoS guarantees is presented and discussed in detail.

Chapter 5 discusses the various issues involved in resource allocation for real time traffic. The existing link level scheduling schemes are examined. A new slot allocation/scheduling policy derived from the Hop-Laxity scheme is presented along with the buffering strategy and the buffer overflow control scheme used in the protocol.

In chapter 6, we discuss the discrete event simulation and the results obtained from the discrete event simulation of the protocol. We also discuss the improvements to the protocols and future work, and finally the conclusions are given.

Chapter 2

ATM Networks

2.1 ATM - Asynchronous Transfer Mode

The need for broadband communications in fixed networks has led to the development of a new switching technology called the Asynchronous Transfer Mode (ATM) [2]. The ATM technology will offer more flexibility and performance than existing technologies. ATM can integrate the transmission of all kinds of data into a single network usually called the Broadband Integrated Services Digital Network (B-ISDN).

The ATM technology is currently developed by two standardization organizations; the ITU-T and the ATM Forum. The ITU-T (International Telecommunications Union - Telecommunication Standardization Sector) acts as the world wide telecommunication standardization body. The ATM Forum, on the other hand, is an operator and manufacturer driven organization established to develop and standardize the ATM technology. ETSI (European Telecommunications Standards Institute) has adopted a policy of producing ATM standards compatible with those of ITU-T.

2.2 ATM and B-ISDN

Traditional telecommunication networks are specialized; there are clear relations between services and networks. Many of the telecommunication services have their own networks and those networks are typically not very suited for supporting well other services than those initially intended to be supported. As an example, the Public Switched Telephone Network (PSTN) offers two-way voice connections while the broadcast of high quality audio programs is done using specialized radio networks. Television signals are distributed using at least three totally different transmission technologies; broadcasting, cable and satellite links. For data communication, the range of technologies used is even more extensive.

Our society is becoming more and more information intensive, and both the number of services and the number of users are expected to grow dramatically. The new services will require higher bit rates per user than the existing networks can offer. It would be ineffective to build a new network for every new service. Therefore, the new technology should also be able to support future services; services that we know nothing about when the technology is developed. The new system should also be able to support all the services provided by the existing specialized networks.

The B-ISDN vision is to support all kinds of services in a single network. To fulfill the vision, B-ISDN needs an extremely flexible switching technology. The ATM technology has been developed to be able to fulfill the needs of the B-ISDN. While ATM may be considered a transfer mode for physical transmission enabling very high data rates, B-ISDN is a network specification exploiting ATM technology.

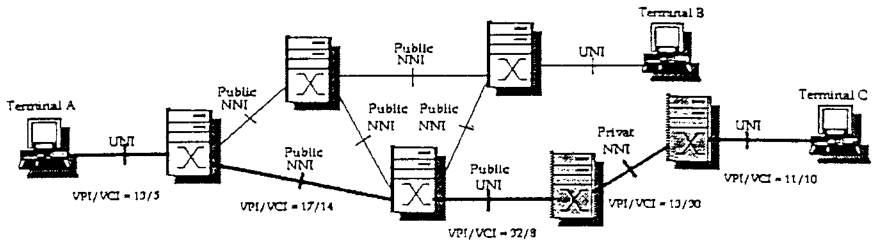

Figure 2.1: ATM terminals A and B have an active virtual connection across an ATM network

2.3 The ATM network

An ATM network consists of ATM switches that are connected to each other using an interface called the Network-Node Interface (NNI). Terminals are connected to the network using User-Network Interface (UNI). If there are privately owned switches in the network, the interface between the public and private part of the network is called the Public UNI and the connection between two private switches a Private NNI (P-NNI). The ATM connection is called a virtual connection (VC). There are two main types of virtual connections; Permanent Virtual Connections (PVC) and Switched Virtual Connections (SVC). Figure 2.1 shows an example of an ATM network, where terminals A and B have an active switched virtual connection. In a switch, the traffic on a virtual connection is routed according to two identifiers; the Virtual Path Identifier (VPI) and the Virtual Channel Identifier (VCI). Besides the VPI/VCI, the identity of the incoming link is also used to identify a VC. A VPI/VCI is allocated for a virtual connection by the switch when the connection is set up, and it remains unchanged for the entire life time of the connection. It should be noted that the VPI/VCI values of a single connection are most likely to be different on different links.

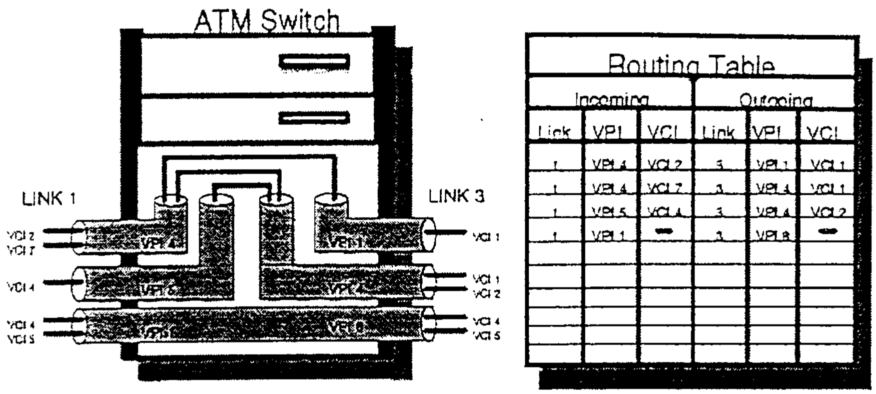

Figure 2.2: The routing of virtual connections in an ATM switch is based on the information of a routing table.

In ATM data is carried in small packets called cells. The cells carry their VPI/VCI identifier in their header and this identifier is used in the switches to route the cell. There is a routing table in each switch with routing information for all active connections passing through the switch. The routing information includes the new VPI/VCI and the new outgoing link for every incoming VC. A virtual path (VP), identified by a VPI, consists of many virtual channels. The routing can also be done only according to the VPI. In this case the routing is called virtual path routing and the value of the VCI remains unchanged. The relation between the routing table, VPI and VCI in a switch is shown in Figure 2.2.

Signaling protocols are needed for establishing and managing connections across a network. The signaling protocols to be used in ATM are currently under development. The two main candidates for the signaling over the UNI are the Q.2931 standardized by the ITU-T and the UNI-signaling of ATM-Forum. The UNI signaling offers the terminal a language to speak with the ATM switch in order to manage the virtual connections. The signaling to be used between the switches of an ATM network is called NNI signaling and it is used for call control and to route the virtual connection across the ATM network.

When a terminal wants to establish a new connection (virtual circuit) with some other terminal, it uses the UNI signaling (e.g. Q.2931 SETUP message) to request the network to setup the connection. The calling terminal will use a set of traffic and Quality of Service (QoS) parameters to describe the traffic it has and the quality of the connections it requires from the network. These parameters are presented later. Upon receiving this request across an UNI interface, the switch will start the call establishment using switch-to-switch NNI-signaling (if both the calling and called terminal are connected to the same switch, only UNI signaling is used). The routing of a connection in public networks is typically based on the E.164 address standard specified by ITU-T, while most private networks use the NSAP addressing format of the ATM Forum. ATM address of the ATM Forum uses the syntax of the OSI NSAP (Network Service Access Point) address. The interworking between networks using E.164 and NSAP addressing is not too difficult.

When a switch receives a request to set up a connection, it will execute a Connection Admission Control (CAC) function to define whether it can accept the connection or not. If accepted, it forwards the request towards the called terminal. If the connection cannot be supported according to the parameters requested by the calling terminal, the switch will reject the connection setup request. The next switch on the way towards the called terminal receives the request, executes a CAC and forwards or rejects the setup request. If the establishment of a connection is rejected by a switch, a previous switch can try to reroute the connection via some other switches. When the setup request reaches the switch of the called terminal, the request is signaled to the called terminal using UNI signaling.

When a virtual connection is set up with specified descriptors, these descriptors make up a traffic contract between the user and the network. The network will guarantee the type and quality of service agreed as long as the traffic conforms to the traffic contract.

It is important that no terminal violating its traffic contract can lower the QoS experienced by other terminals. It is necessary to ensure that a terminal is not sending more traffic than it is allowed to send. The function implementing this policing is called Usage Parameter Control (UPC). The Generic Cell Rate Algorithm (GCRA) can be used to define conformance with respect to the traffic contract. For each cell arrival, the GCRA determines whether the cell conforms to the traffic contract of the connection. The GCRA algorithm is sometimes called the continuous-state leaky bucket algorithm.

2.4 Traffic over ATM network

ATM is designed to support all kinds of telecommunication services and therefore it should also be able to carry all types of data. This section will present different kinds of traffic and the way they are carried over ATM. A set of traffic classes is defined to support the transmission of different kind of traffic.

2.4.1 Variable bandwidth traffic

All communication is bursty. For some computing applications the required peak bandwidth can be many hundred times the average bandwidth. Many

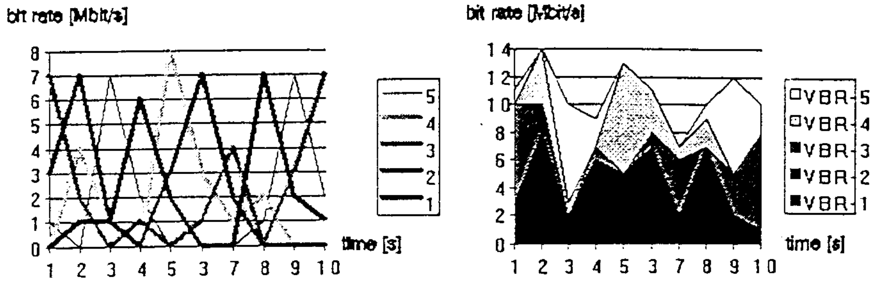

Figure 2.3: Statistical multiplexing of variable bit rate ATM connection. The bit rate of individual connections and the combined bit rate on the link are shown.

of the existing technologies reserve a constant bandwidth for a connection and usually there is only a few bandwidths to choose from. This leads to ineffective use of resources and reduces available bandwidth per connection. The ATM technology is more flexible. The available bandwidth can be allocated to the applications according to their needs. If the variable bandwidth data is time critical it is carried in ATM using the Variable Bit Rate, real-time (rt-VBR) traffic class.

In ATM many connections are multiplexed dynamically on a single link. This introduces a multiplexing effect. Figure 2.3 shows the multiplexing effect when five variable bit rate connections are multiplexed on one link. Each connection has a peak bandwidth of 8 Mbit/s and an average bandwidth of 2.2 Mbit/s. If the peak bandwidth would be allocated to each of the connections, a total of 40 Mbit/s would be needed. In the example only 20 Mbit/s is needed because of the multiplexing effect. The needed bandwidth is calculated statistically and there is always the possibility that the connections would temporarily need more than the total bandwidth available. In this case some of the traffic needs to be buffered and additional delay is introduced. If there is not enough buffer space some data will be lost.

2.4.2 Constant bandwidth traffic

Even if all communication is bursty all traffic is not bursty; coding methods that produce non bursty data are often used. This is the case for example in voice communications using Pulse Code Modulation (PCM), where the voice in sampled using a constant sampling frequency and every sample is coded using a constant number of bits. This is not the most effective coding, but it might be the cheapest.

Constant bandwidth traffic needs to be carried across a network the way it is; at a constant bit rate and with a constant time between samples. Many existing networks offer only constant bandwidth connection. So, a new network should support the transmission of constant bandwidth traffic in order to be compatible with older systems. In ATM it is possible to reserve constant bandwidth for a connection using the Constant Bit Rate (CBR) traffic class. When older networks typically support only one bandwidth, the bandwidth of an ATM CBR connection can be defined when the connection is established.

2.4.3 Non-real-time traffic

All traffic can be divided into real-time and non-real-time traffic (sometimes the division is not clear). Real-time data has to be transmitted with a certain delay and delay variation, otherwise it will lose its value to the recipient. Interactive services like telephony and video-phony are typical examples of services that produce real-time data; a telephone produces typically CBR and a video-phone rt-VBR traffic.

There is also non-real-time data. Computer data is typically non real-time data. Non-real-time data has no critical delay requirement, so it can be buffered more in the network. This introduces the second multiplexing benefit of ATM; non real-time data can be transmitted between the peaks of the real-time data. In ATM there are three traffic classes used to carry non real-time traffic; the Variable Bit Rate non real-time (nrt-VBR) class, the Available Bit Rate (ABR) class and the Unspecified Bit Rate (UBR) class.

2.4.4 ATM traffic classes and their parameters

The ATM forum has defined a set of traffic classes and their parameters. The traffic classes are called service categories. The ATM traffic classes are summarized below.

- **Constant Bit Rate (CBR):** The CBR connection type is used to carry constant bit rate traffic with a fixed timing relationship between data samples.

- **Variable Bit Rate - real time (rt-VBR):** The rt-VBR service class is used for connections that carry variable bit rate traffic, in which there is a fixed timing relationship between samples.

- **Variable Bit Rate - non-real time (nrt-VBR):** The nrt-VBR service class is used for connections that carry variable bit rate traffic in which there is no timing relationship between data samples, but a guarantee for the mean transfer delay and cell loss is required.

- **Available Bit Rate (ABR):** ABR is non real-time traffic class like the nrt-VBR, but the bandwidth allocation is based on the "best effort" approach. The network will give the connection as much bandwidth as possible, but it has to meet the minimum values of the traffic contract. The

ABR introduces also flow control per virtual connection and it is best suited for computer data transfer.

- **Unspecified Bit Rate (UBR):** The UBR service does not offer any service guarantees. The user is free to send any amount of data up to a specified maximum while the network makes no guarantees at all on the cell loss rate, delay or delay variation that might be experienced.

The traffic to be sent is defined by a set of traffic parameters:

Peak Cell Rate (PCR), Sustainable Cell Rate (SCR), Burst Tolerance (BT), Minimum Cell Rate (MCR), Cell Delay Variation Tolerance (CDVT). The PCR determines the maximum and the MCR the minimum cell rate of the traffic. Long-term average cell rate and the size of the maximum burst of contiguous cells that can be transmitted is determined by the SCR and BT together. The CDVT is used to prevent the delay jitter caused by network equipment between the terminal and the ATM switch offering the UNI signaling from effecting the UPC function.

2.5 The protocol stack of ATM

The ATM standards contain two layers; the ATM layer and the ATM Adaptation layer (AAL). The ATM protocol reference model can be seen in Figure 2.4.

The ATM layer somewhat corresponds to the OSI layer 2 (Data Link Layer - DLL). The ATM layer transports all information in ATM cells that are only 53 bytes long. On top of the ATM layer is the AAL. Its function is to offer an interface for the upper layer to use. The ATM can be used on top of different physical layers (PL) and media; multi-mode and single-mode fiber as well as coax

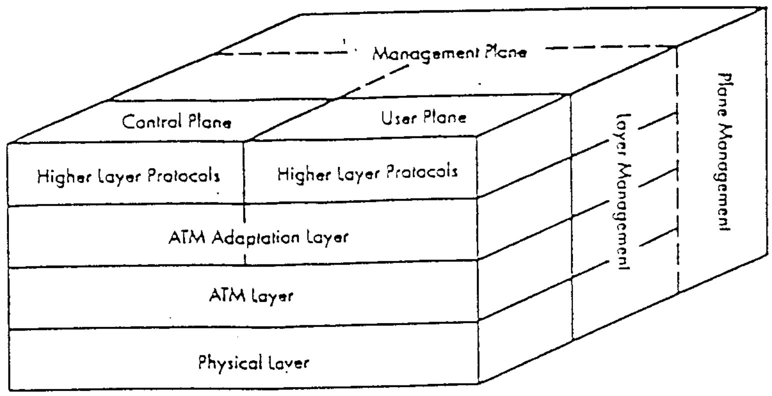

Figure 2.4: The ATM protocol reference model

and even twisted pair can be used. The management plane controls user plane, control plane and the layers in them. The protocol layers of the user-plane in different nodes of an ATM network are shown in Figure 2.5.

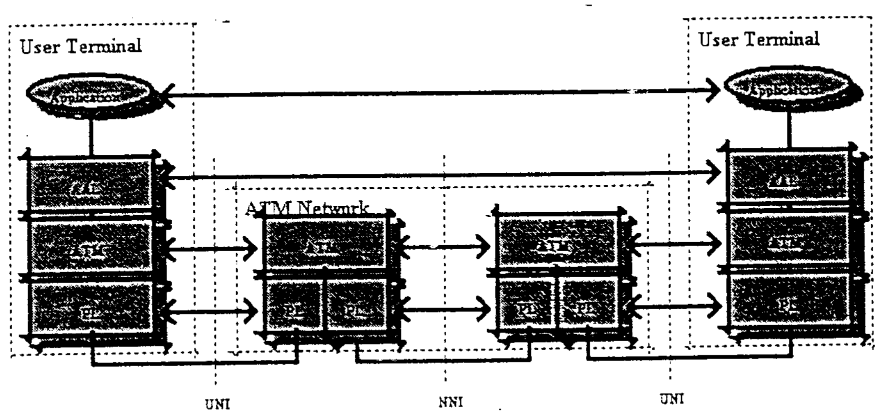

Figure 2.5: The protocol layers of an ATM network

2.5.1 Physical layer

Many different physical layers can be used in ATM. The most common ones are the Synchronous Digital Hierarchy (SDH) and Synchronous Optical NETwork (SONET). These two standards, one developed in Europe and one in America are quite similar; the European SDH is based on the American SONET. They can use both multi-mode and single mode optical fiber as the transmission medium. Also coaxial cable and twisted pair can be used in ATM, but the bit rates achieved are lower than those achieved on fiber.

There is one architectural difference between the physical layer of ATM and the OSI physical layer; the ATM physical layer comprises of two sub-layers. These are the physical medium (PM) and the transmission convergence (TC) sub-layers. The PM sub-layer is quite close to the OSI physical layer; it takes care of the transmission of the bits across the physical medium. The TC is more ATM dependent and deals with the functions necessary to support the ATM cell type data transfer. These functions are for example cell delineation, error control and insertion of empty cells during idle periods to keep bit synchronization at the PM sub-layer.

2.5.2 ATM layer

The flexibility of an ATM network is mainly achieved though the use of small data packets called ATM cells. These cells are routed through the network on the ATM layer. The ATM cell is 53 bytes and contains a 5 byte header. The header is modified by every switch on the virtual connection, but the payload remains the same across the network. The structure of the ATM cell is shown is Figure 2.6 [3].

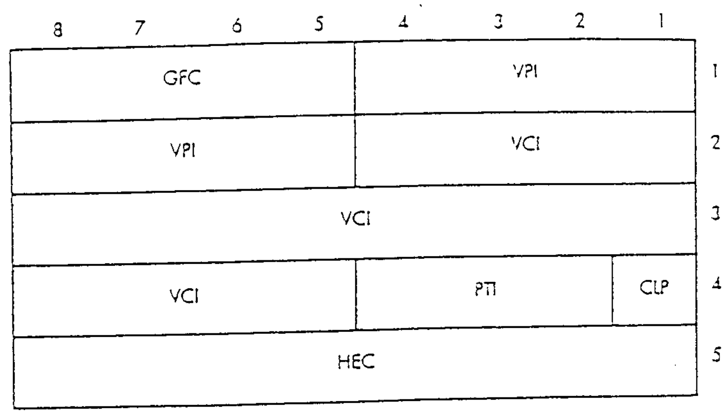

Figure 2.6: The fields of the ATM cell at the User-Network Interface

The structure of the ATM cell is similar in different nodes of the network with one exception. The only difference is the use of the four last bits of the first byte; in UNI the bits are used for the GFC field while in NNI they are used to extend the numbering space of the VPI. The use of the GFC is not specified yet. The VPI can be 8 or 12 bits depending on the interface, while the VCI is always 16 bits. Some of the combinations of VPI and VCI are used for special purposes, like the VPI/VCI 0/5 used for point-to-point signaling at the UNI.

The three bits of PT indicate the type of information carried in the cell. The content of the PT is called the payload type identifier (PTI) and its values are shown in table 2. The third bit of the PTI indicates whether the cell is a data cell (0) or a management cell (1). In data cells the second bit indicates whether congestion was experienced (1) or not (0).

2.5.3 ATM Adaptation layer

The ATM Adaptation layer (AAL) adapts different services to the ATM network. The AAL receives the upper layer PDU and fits it into one or more ATM cells. It also provides timing and error control services. There are four different AALs each intended to support certain services. A special AAL to be used for signaling, the signaling AAL (SAAL), is also specified.

The AAL1 is intended for the transmission of constant bit rate (CBR) traffic. To support the CBR services and circuit emulation, functions for source clock recovery and control of cell delay variation are included. The cell delay variation needs to be taken into account, to avoid interruptions in the bit stream because of a delayed cell.

The AAL2 is expected to be used mainly for variable bit rate communications.

The ATM adaptation layers 3 and 4 have been merged into the AAL3/4. The AAL3/4 performs error control on two levels; AAL-PDU and ATM cell. Error control is added both before and after the user data in the PDU. In addition to that, a cyclic redundancy code (CRC) field is added to every cell together with a multiplexing identification (MID) field that can be used to multiplex many connections onto a single VC.

The AAL number 5 is best suited for data communications. A trailer containing a CRC (used for error control) is inserted in the last cell of each PDU. The last cell of a PDU is indicated using the first bit of the payload type identifier (PTI) of the header of the cell. For the time being AAL 1 and AAL 5 are the most common AALs used.

The SAAL provides reliable transport of signaling messages in an ATM network. The SAAL consists of service specific parts implemented on top of an

AAL5. The new functions include the retransmission of lost signaling messages.

2.6 ABR flow control

The wireless ATM system presented in this thesis is supposed to carry mainly ABR traffic. In the ABR service the source adapts its rate to changing network conditions. The key feature of the ABR traffic class is its flow control method that allows it to dynamically use the free link capacity. The flow control makes ABR well suited for the transmission of cell loss sensitive data. The ATM Forum decided the flow control method to be rate based, where the transmission rate of the source is controlled. The other candidate was the credit based system, where the traffic is controlled per link according to the number of buffers available in the next switch.

There are two different flow control methods in ABR; binary feedback and explicit feedback. The difference between them is how the transmission rate of the source is controlled. In the binary feedback method the source is requested to decrease its transmission rate, but no new cell rate is indicated. The size of the decrement has been agreed at the connection set up. The source can also be prevented from increasing its transmission rate. The source can increase its transmission rate with a specified increment when allowed. In the explicit feedback method the network gives the source the explicit maximum transmission rate it can use [4].

In the Explicit rate (ER) feedback method, ER is used to limit the source rate to a specific value. At the start of the service the source rate is set to a requested rate (e.g., PCR) by the source and may be reduced by any node in the network. This rate is known as CCR - Current Cell Rate. MCR - Minimum Cell Rate, indicates the minimum cell rate of the connection. The allowed cell rate (ACR) is the cell rate the network dynamically allows the source to use depending on the amount of available bandwidth on the connection. The PCR traffic parameter determines the maximum value of the ACR, while the MCR determines the minimum value.

The explicit feedback is more powerful because it adjusts the transmission rate of the source faster than the binary method. On the other hand, the implementation of a binary feedback method is likely to require less computation resource in the switch than the explicit method.

2.7 Conclusions

The ATM technology has been presented in this chapter. The aim was to give the user an overview of ATM and to present more closely some of its features that will have special importance from the point of view of this thesis.

The ATM cells are routed in a switch according to a routing table of the switch and the connection identifiers in the cell. The routing table can be modified by a signaling application in the switch, so modification of an ATM route is possible without involving terminals into the process. The transportation of upper layer traffic in many small cells that cannot be identified means that the whole PDU is likely to be lost, if a cell is lost or the cells are received out of order. This places hard requirements on a mobile system where the sequence of the cells is not naturally maintained and where the communication links can be temporarily down due to various reasons. The statistical multiplexing principle is very important in a mobile system, where the radio interface is an expensive resource. The ABR traffic class and its flow control have an essential role in the WATM model presented in this thesis.

Chapter 3

Wireless ATM

3.1 What is Wireless ATM ?

Wireless ATM is essentially the wireless extension of ATM networks as shown in Figure 3.1. Wireless ATM is mainly considered as an "access to an ATM network" issue. Depending on what kind of ATM network is to be accessed, different aspects of wireless networks need to be addressed. In the LAN scenario, wireless ATM can be considered as an extension of LAN for mobile users. In ATM to the home scenario, wireless ATM has the appeal of extending the set top box into portable appliances such as television receivers. More important, it has the potential to address a need to provide residential access for alternate carriers after deregulation of the local communication markets.

For a Personal Communication Network (PCN) scenario, the WATM system consists of a hierarchical ATM switching network for interconnection of PCN microcells, each of which is serviced by a high-speed, shared access radio links based on ATM-compatible cell relay principles. In this approach, an ATM cell serves as the basic unit for protocol processing and switching in both wired and

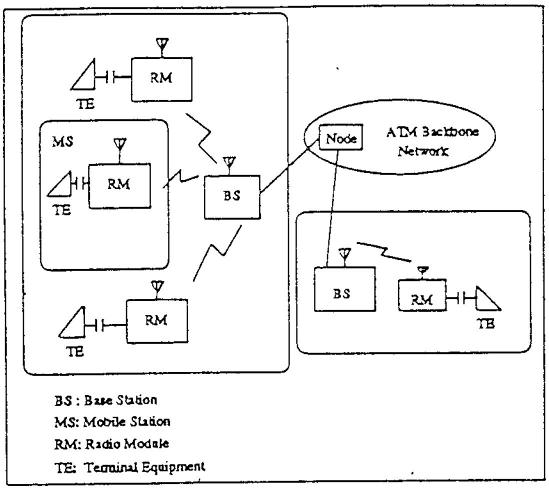

Figure 3.1: Concept of Wireless ATM

wireless portions of the network. The overall transport architecture is based on the ATM protocol stack, with appropriate extensions where necessary to support new mobility and wireless channel specific requirements. In particular, the wireless segment of the network will require additional medium access control (MAC) and data link layers for channel sharing and error control on the radio links. If properly designed, these wireless channel specific layers should be able to provide a degree of transparency for a useful subset of broadband/ATM services.

3.2 Why Wireless ATM ?

A typical reaction to the concept of wireless ATM is to question the compatibility of several aspects of ATM technology to the wireless channel. First, considering the fact that ATM was designed for media whose bit error rates are very low (about 10^{-10}), it is questioned whether ATM will work at all in the highly noisy wireless environment. The environment in question is typically a multi-access channel that may also be time varying. Second, the wireless channel is an expensive resource in terms of bandwidth, whereas ATM was designed for bandwidth rich environments. ATM effectively trades off bandwidth for simplicity in switching. However, if these problems are solved there are significant advantages as the following [5] :

- In order to avoid a serious mismatch between future wireline and wireless network, it is important to begin consideration of broadband wireless systems with similar service capabilities as the broadband terrestrial networks. Personal communication networks introduced into the future multimedia application scenario should provide the service features listed in the previous section for the multimedia applications. Clearly, implementation of these broadband features on the wireless medium is a more difficult technical challenge than for fiber, but it is important to aim at system designs, which provide qualitatively similar attributes.
- The availability of qualitatively equivalent ATM service classes in a seamless manner would provide significant benefits of uniformity to network operators, service providers and terminal designers.
- The residential delivery of application such as Video on Demand etc, WATM is a useful way of providing broadband access without investing in new infrastructure.

3.3 The WATM usage scenario

This section will present the vision behind the WATM model. The scenario presents the users, who are foreseen, to be the first ones requiring wireless access to ATM networks. Also the terminals and applications the users are likely to use are discussed.

It is expected that the need for wireless ATM services will first materialize in large companies having their own networks in their premises. Fixed ATM based local area networks are expected to replace local area networks currently used by these companies. Later on, other communications, such as fax, telephone and video-conferencing, could also be moved to the ATM network. Users might request for the ability to move while using the LAN, thus generating a demand for a wireless access to the ATM network. The system envisioned is a wireless customer premises ATM network supporting both fixed and mobile terminals.

A typical mobile terminal in the WATM usage scenario is a laptop computer used by an office worker. The user wants to be able to use his or her computer in different parts of the office building and also in other frequently visited buildings. The network services should be equally available in office rooms, in meeting rooms, in the cafeteria, etc. When the user works at his own desk, he should be able to use the fixed access available to get greater performance than the wireless access to the network can offer. Being able to used both fixed and wireless system will also lower the overall load of the wireless system.

Another likely usage scenario is that of residential environment [6]. Here the WATM network would consist of several appliances, for example, PC laptop, printers/fax machines, home appliances, digital VCR etc. Important considerations for such a system include low cost, plug and play operation, and high flexibility in terms of setting up partial networks and changing network configuration over time. It is expected that these devices will have a low degree of mobility. High data rates of up to 25 Mbps are expected over this network.

The applications foreseen to be used are typical B-ISDN applications that must be supported for mobile users with an acceptable QoS. The QoS expected from the wireless ATM system could be somewhat lower than the QoS of a fixed ATM system. The user is assumed to realize that a small loss in QoS is the price paid for the mobility gained. For example, the Cell Loss Rate might be larger (resulting in somewhat lower "goodput"), and a short interruption in the connection because of a handover is tolerated. For non-real time connections the aim should be a lossless handover, that is, some delay is inflicted but no data is lost. Also the security issue has to be considered as in any wireless system.

Some possible applications in the WATM usage scenario are:

Computing: Typical computing applications foreseen to be used in the proposed system are client server applications, file systems, e-mail delivery, fax, group-ware and computer games. A powerful computer in the network could assist the mobile terminal to run computation intensive applications like Computer Aided Design (CAD). Connection quality close to the one offered by a fixed ATM LAN is needed. Increased CLR and interruptions will increase the retransmission frequency, so the delay performance is expected to be weaker.

Audio: The audio applications foreseen include public announcement, high quality telephone or a wireless equipment for Digital Audio Broadcasting (DAB) quality program production. The ordinary telephone service is the most commonly used telecommunication service and is likely to be requested by the user also in the future. It is possible that more than one simultaneous connection exists for a terminal.

Video-phone: Powerful lap top computers likely to available in the future are good platforms for video-phone applications. While the need for residential video-phony is unlikely to materialize, the demand for a video-phone in office environment has been foreseen. The picture quality does not need to be excellent, because of the limited display facilities in the terminal (limited resolution of 10-12 inch displays of lap top computers).

3.4 System Design Criteria

In addition to the multi-service requirements indicated in the preceding sections, many other factors need to be taken into consideration in the design of the next generation PCN. Some of the major system level design objectives [5] can be summarized as follows:

- Flexible multi-service capability including voice, video, data and multimedia.
- Good quality-of-service QoS for various service types.
- High degree of compatibility with future broadband networks, including ATM/B-ISDN.
- Low terminal cost/complexity/power consumption.
- High net spatial bandwidth efficiency (i.e. $bps/Hz/Km^2$).
- Efficient, scalable, and moderate cost network architecture (including the fixed portion).
- Compliance with the regulatory constraints.

3.5 System Architecture

A typical architecture for a mobile communications network consists of base stations (BSs) servicing a set of mobiles within a small coverage area called a cell. Thus, a BS oriented network imposes structure on the network, thereby restricting flexibility and may not address the plug and play requirement of the consumer.

An adhoc network architecture is one where all network devices are connected in a random fashion and is marked by the absence of a central controller (BS). However complicated protocols have been proposed to support an adhoc network topology, thereby increasing the cost to the consumer. Thus it is not clear which approach should be chosen. In our system the BS oriented approach will be pursued because of the simplicity of implementation of such systems, as compared to the adhoc architectures.

3.5.1 Centralized Network Architecture

A base station (BS) oriented network architecture is essentially a centralized architecture consisting of a number of base stations (BS) and/or a mobile switching center (MSC) through which communications between Wireless Terminals (WTs) are made. Both "control and management" and data transmissions are done in centralized nodes. That means the control and management messages from wireless terminals must be sent to the BS, where control actions are determined and transmitted back to the wireless terminals in a broadcast manner. Meanwhile, transmission of data packets between two WTs must be done through the BS as well.

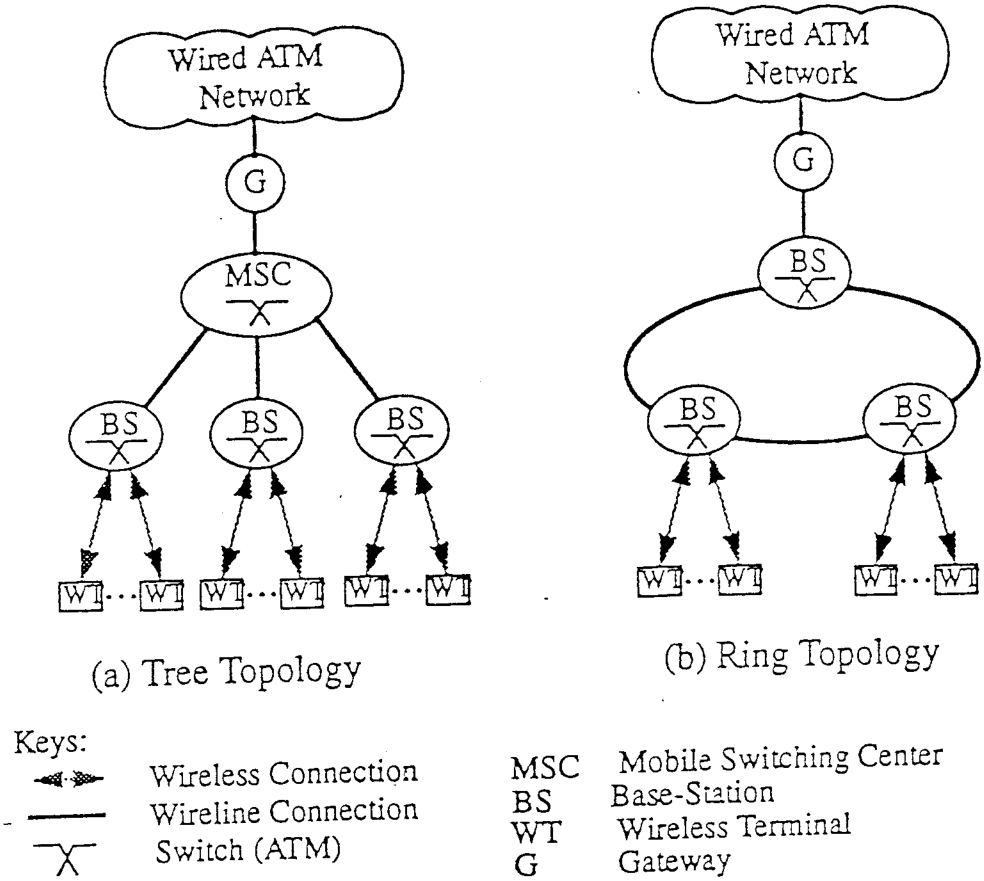

Figure 3.2: Centralized BS Architectures

The concept of centralized architecture has been used for years in cellular communications, e.g. GSM, TACS, etc. Service areas in this case are divided into small regions called "cells", each of which is served by a BS. Communications between the WTs are done via the BSs and/or a MSC. In general there are two types of centralized architectures, namely the tree topology and the ring topology, as shown in Figure 3.2.

In tree topology, the switching is performed in a hierarchical manner While switching for intra-cell calls is done using the BS, switching for inter-cell calls is conducted in the MSC. Normally the switch in the MSC is more sophisticated and complex than those in the BSs.

In the ring topology, the switching is performed in a "distributed" way. While switching for intra-cell calls is done in the BS as in the tree topology, switching for inter-cell calls is conducted by passing the call to the destined BS switch around the ring. Unlike the tree topology, the ring topology uses identical BSs and switches which makes it very scalable. This topology has been proposed in [7] for ATM local area networks.

3.6 Protocol Layering

The design of the WATM system should follow a protocol layering that is harmonized with the ATM protocol stack - this is important to ensure transparency in the design of the system. The approach taken here is the "native-mode" ATM approach. The wireless channel specific physical, medium access control and data link layers are below the ATM network layer as shown in Figure 3.3. The ATM layer can then be designed independent of whether a wired or a wireless MAC/PHY layer is used. This is highly desirable since this makes the design of MAC protocols for the wireless subsystem future-proof.

Based on the native mode ATM approach, a generic layered model of wireless-ATM has been proposed (Figure 3.3). This model can be applied to a BS oriented wireless ATM systems as shown in Figure 3.4. This basically follows the approach suggested in [8]. At the base station, ATM switching is performed and resource and mobility management are handled using separate MAC and PHY, the BS can handle both wired and wireless traffic as shown by the shaded application

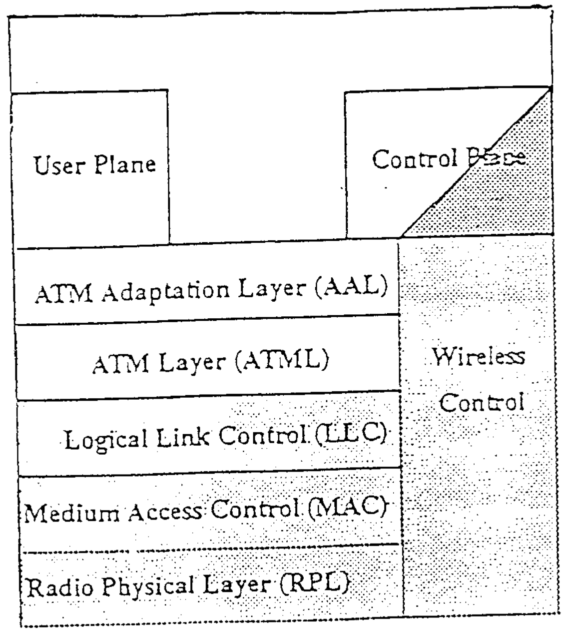

Figure 3.3: Protocol Stack for a WATM network

path between a wired ATM terminal and a wireless ATM terminal in Figure 3.4.

Note that the ATM layer is concerned only with the data channel and not with the control and the management channel. The wireless MAC layer essentially specifies how the wireless terminals interact with the control and management channel. Once the reservation is made for a particular amount of channel bandwidth, ATM cells from the output of the ATM layer are sent over the wireless channel. Also note that the "wireless control and management function" resides inside the "resource management and mobility support" block. Although it is not shown in the figure, the "ATM control and management" function interworks with the wireless control and management functions, which in turn, can communicate with the wireless MAC/PHY layers. This information exchange is required to setup and tear-down wireless physical connections.

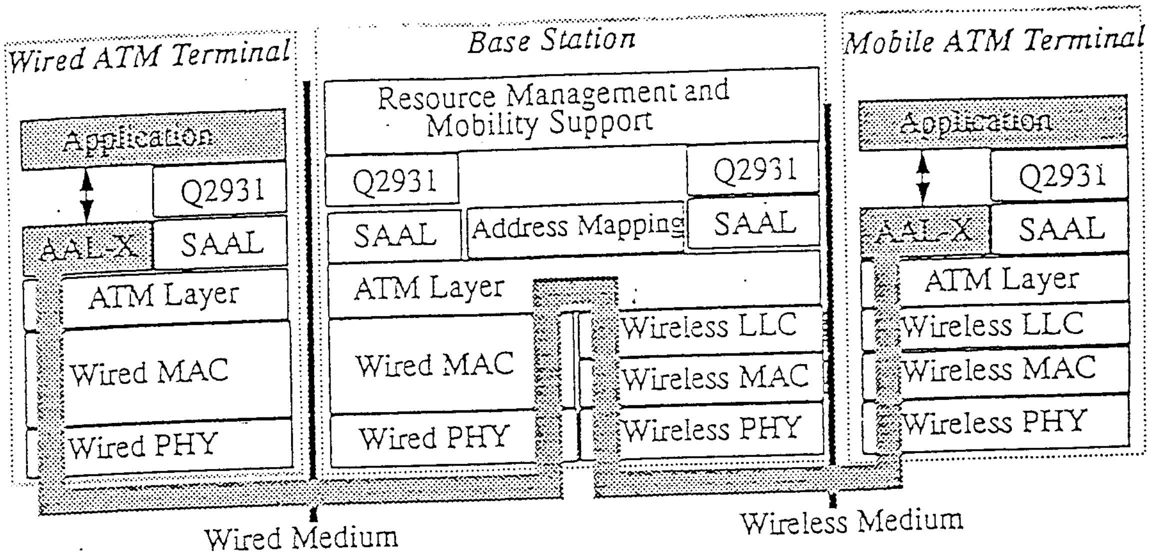

Figure 3.4: Wired/Wireless ATM layered model for a BS oriented system

Chapter 4

MAC Protocols

4.1 Medium Access Control (MAC)

A major issue in the WATM system design is the selection of a suitable channel sharing/media access control (MAC) technique at the data link layer. The MAC technique used will have a significant impact on user performance, system capacity and remote terminal complexity, and is therefore an important design parameter. As discussed earlier the WATM system will be required to handle a diverse mix of traffic types. Thus the adopted MAC approach must provide mechanisms to deal with each of these B-ISDN service types at reasonable quality-of-service (QoS) levels. The design challenge thus is to identify a wireless "multimedia capable" MAC approach that provides a sufficient degree of transparency for many ATM applications.

4.2 Types of MAC protocols

Medium access or multi-access (MAC) protocols attempt to efficiently and equitably allocate use of a communications channel to independent, competing users.

Errors result when two or more users simultaneously (in time and frequency) attempt to transmit over a channel.

When such a conflict occurs, all involved packets must be retransmitted. A user hears if its transmission was successful, one round-trip propagation time after transmitting a packet over the multi-access channel. Propagation delays across the channel range from millisecond values for LANs to approximately 0.25 s in satellite applications.

Many channel access protocols have been proposed and analyzed in the past few decades [9]. Each scheme has its advantages and limitations, providing acceptable performance only in certain environments and with certain types of channel traffic. Protocols are grouped into five classes [10] :

- Fixed assignment
- Random access
- Centrally controlled demand assignment (reservation based)
- Demand assignment with distributed control
- Adaptive strategies and mixed modes

In addition three network environments are characterized : satellite, ground radio and local area. Each environment has its own unique characteristics, such as magnitude of propagation delays. The performance of the protocol depends a large extent on the environment.

Fixed-assignment techniques, such as TDMA and frequency division multiple access (FDMA), incorporate permanent subchannel assignments (in time-frequency domain) for individual users. These "classical" schemes perform well with stream-type traffic (each user transmitting a steady flow of messages), such as voice. At all times, a large percentage of subchannels carry user traffic. The result is high utilization of the communications channel, and low user response times. Fixed assignment techniques, however, are inefficient in bursty traffic applications and lead to very low channel occupancy rates.

Bursty traffic is serviced more efficiently by random access protocols. The ALOHA and carrier sense multiple access (CSMA) are typical examples. Random access techniques make the full channel capacity available to users, for short periods of time on a random basis. They are packet oriented, whereas the fixed assignment techniques are channel oriented. They dynamically allocate transmission capacity on a per packet basis.

The simplest random access protocol, pure ALOHA, permits users to transmit at will. Whenever one user's transmission overlaps any other user's transmission, a collision occurs, and both messages must be retransmitted. When the channel is lightly loaded, few collisions occur in ALOHA based schemes. Consequently, the expected delay, from arrival of a packet until its successful transmission is quite small. However, these schemes perform very poorly under heavy traffic and are inherently unstable.

For LANs, random access protocols take advantage of the short propagation delays between users. In CSMA, transmission is delayed until the channel is sensed idle. This reduces the number of collisions and leads to high throughput and low delay. If users have the ability to detect collisions (e.g. CSMA with collision detection CSMA/CD), additional performance gains result. Despite the improvements achieved with carrier sense techniques, stability problems persist. More importantly its very difficult to sense collisions in a wireless channel making such algorithms unsuitable for use in wireless environments.

Like random access protocols mentioned above, demand assignment techniques provide channel capacity to users on a demand basis. Unlike random access however, demand assignment involves two stages : a reservation stage followed by a transmission stage. Demand assigned (reservation based) protocols achieve higher channel throughput by requiring users to reserve bandwidth. A portion of channel capacity is required in this reservation stage. The reservation subchannel can be accessed by users either using a multiple access technique - either TDMA or ALOHA type protocol. Short reservation packets are sent to request channel time, the shorter they are, the less the overhead required for reservation. Once channel time is reserved, information packets are transmitted conflict free. Conflicts may occur only on the small capacity reservation subchannel. At low throughputs, though, the message delay is increased over that of random access techniques, however the system is stable at higher loads and has a much better throughput and delay performance.

Control of the reservation and transmission stages can either be centralized or distributed (adhoc). As discussed earlier, for plug and play requirement a distributed architecture is preferable, however this requires complicated protocols resulting in a substantial increase in the system cost.

Many MAC protocols have been proposed for Wireless LAN/ATM systems [11-17]. These protocols are based on a wide variety of medium access strategies ranging from slotted ALOHA [11] to reservation [15-17] and polling based protocols [12-14]. Notable among these are some polling and reservation based protocols. Some of these are reviewed in the following section.

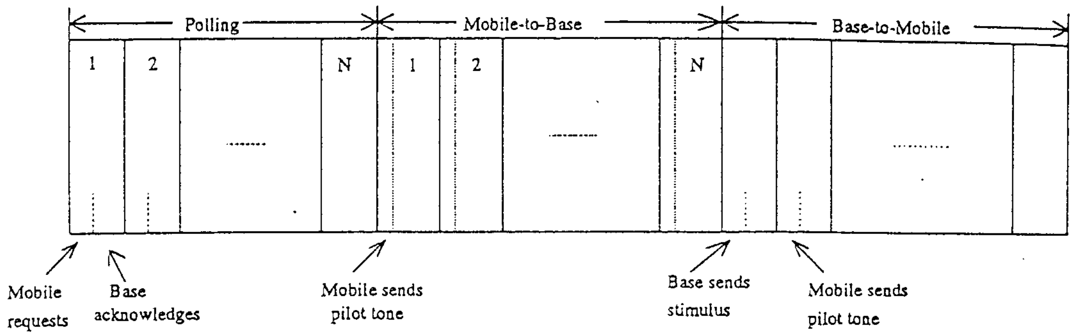

Figure 4.1: A Polling Cycle

4.3 Some existing MAC protocols

4.3.1 Polling based protocols

Acampora et al [14] propose a polling based protocol which uses a virtual connection tree topology. The tree contains its root node, branches and leaves (the base stations). The footprint of the tree is defined to be the collection of radio cells spanned by the tree. Within each radio cell, the entire bandwidth of the radio channel is available to each accessing mobile ; that is each mobile sends and receives ATM cells to the base station using the entire system bandwidth. The base station polls the mobiles and if the mobiles have packets to send they do so in response to the polling. This is done for each mobile registered with the base station in the cell.

A mobile cannot send ATM cells until it receives a token. Any mobile in receipt of a token must either send a sequence of ATM cells (up to a maximum limit) or a short pilot tone. Within each polling cycle the base station sequentially sends a token to each remote and immediately prepares for a reply from that remote. In the downlink period the base station sequentially transmits

ATM cells to each mobile if there are any packets destined for the mobile. This process repeats until the base station has no cells to send or a time-out interval occurs, the time-out interval, being defined as the maximum allowable period of time between the receipt of a signal from each remote, has elapsed. The time-out interval reflects the fact that the environment is continually changing and, if too long period of time has elapsed it might be too difficult to adapt to the physical link (update the set of antenna weights). At this point the base station must repoll each of the mobiles within its cell. The efficiency for the above protocol is defined as:

$\% = (T - NP)/T$, where

T: The time-out interval,

N: The number of mobiles

P: The polling time for each mobile.

Thus the efficiency decreases linearly with the number of mobiles to be polled within the cell.

- Chen et al [12-13] present polling based protocol called Random Addressed Polling and Group Random Addressed Polling (RAP and G-RAP). The Random Addressed Polling (RAP) protocol is a contention based protocol which limits the number of simultaneous transmission attempts through a random selection. The protocol separates the up and downlink channels, and is mainly concerned with the uplink. The base station transmits a ready message when it is ready to receive uplink messages. In response each active node (in total N active nodes) will choose and transmit a random number (0,1,...,p-1) using Code Division Multiple Access (CDMA) which allows the base station to receive all the transmitted messages simultaneously. The base station will detect N*($\leq N$)

different numbers ; for each number the base station will broadcast this number and the terminals which chose this number will transmit a packet. If several terminals chose the same number, collision will occur. The base station will provide feedback (ACK/NACK) and for the unsuccessful transmissions the procedure is repeated.

The Group Randomly Addressed Polling (GRAP) improves upon RAP by introducing a superframe structure consisting of p+1 frames, named G0,G1,...,Gp. Each frame is a separate RAP frame. A new node will frame Gp, whereas a node which in the previous superframe was successfully polled with random number i (and has more packets to transmit) will join frame Gi in the current superframe.

Note that in the above protocols there is no notion of priority traffic and also an impatient customer may not be able to transmit if collisions occur repeatedly, hence these in our view are not very suitable for making QoS guarantees for priority traffic.

4.3.2 Reservation based protocols

PRMA

In Packet Reservation Multiple Access (PRMA) [15], time slots are grouped in frames and each time slot is recognized as "reserved" or "available" according to an acknowledgment message from the base at the end of each slot. When a mobile with periodic source successfully transmits a packet in an available slot to the base station (using slotted ALOHA protocol), the mobile reserves that slot in future frames and there are no subsequent collisions with packets from other terminals. At the end of each burst, the mobile releases its reservation by leaving the reserved slot empty. "Random information packets" also contend for available time slots using slotted ALOHA. However, when a random packet is successfully transmitted, the mobile does not obtain a time slot reservation.

Clearly at high loads the performance of this protocol suffers. Also this protocol though supports voice traffic but is not suited for VBR traffic because of the bursty nature of the traffic along with the fact that certain kinds of VBR traffic are delay sensitive. Moreover PRMA wastes one time slot at the end of each burst transmission. This leads to inefficient utilization of the channel especially at low loads. Again there is no notion of priority among various packets.

DQRUMA

Distributed queuing request update multiple access (DQRUMA) is a demand assignment channel access protocol. Karol et al [16] consider a cell in a wireless network with a base station and N buffered mobiles. ATM cells arrive at the mobiles, according to in general, a bursty random process. The packets are buffered at the mobile until they are transmitted "uplink" to the base station (according to the channel access protocol). The base station broadcasts packets "downlink" destined for the mobiles within its cell. The uplink (mobile-to-base) and downlink (base-to-mobile) are physically separate i.e. on different frequency channels. Figure 4.2 shows a flow-chart of the DQRUMA protocol. There are two kinds of time channels - the Request Access (RA) and the Packet Transmission (Xmt) channels are formed on a slot by slot basis. As the base station receives transmission requests (by listening to the uplink RA channel) from the mobiles, the base station updates the appropriate entries in a *Request Table*. The request table contains an entry for all N mobiles in the system. When a packet arrives

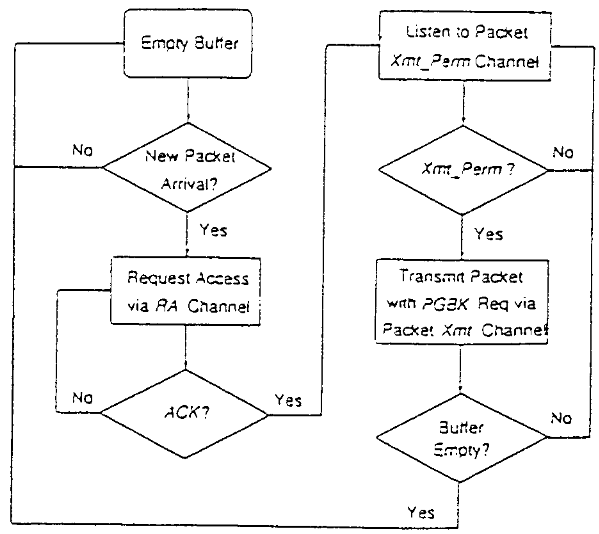

Figure 4.2: DQRUMA protocol

at a mobile with its buffer empty, the mobile sends a *Xmt_Req* to the base station via the uplink RA channel - the RA channels are based on a slotted ALOHA protocol. If the base station successfully receives the request from the mobile (no collisions occur) it sets the corresponding entry in the Request Table to indicate that the mobile has a packet to transmit. The base station also acknowledges reception of the Xmt_Req by broadcasting the mobiles ID. Once the mobile receives positive acknowledgment, it listens to the downlink channel for permission to transmit. The transmit permission is given each time slot to one of the mobiles that has a non-empty Xmt_Req field in the Request Table.

Each time slot that the mobile transmits a packet it also includes a 1 bit piggybacking request if it has more packets to transmit. This serves as a contention free Xmt_Req.

Note in this protocol detailed state information is not being sent to the base station - the base station only knows if a mobile has any packets to transmit or not and thus cannot give priority to one mobile over another if required. Also for highly bursty traffic the Xmt_Req messages have to be sent in contention mode.

4.4 Reservation Based MAC Protocol

In our WATM system we design a reservation based MAC protocol for a centralized network architecture. This section describes a method of establishing and maintaining a connection (accessing the channel), between two entities over a wireless channel. This scheme assumes that all WTs must first register into a WATM network by using some of the management functions of the MAC layer.

4.4.1 Basic Concept

The basic idea behind the reservation-based MAC protocol is a sequence of Control-Data Frames (CDF); each of which consists of a control phase followed by a data phase [18] as shown in Figure 4.3. During the control phase, wireless terminals who wish to transmit in the following CDF send a message to the base station (central controller) which allots (reserves) the slots among different users. During the data phase, the wireless terminals can then transmit in the slots allotted to them. For the purpose of illustration, we consider only the time axis where time is divided into slots; each of which equals the length of a control packet or data packet plus some guard time. The same idea can be easily extended to a more complicated time-frequency plane as well.

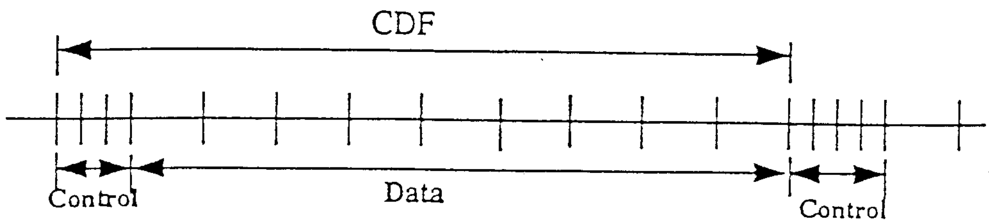

Figure 4.3: Control Data Frame

We allow the number of data slots in a CDF to be variable up to a maximum number. This is a design parameter for the protocol and needs to be determined with the help of a discrete event simulation of the protocol.

In ATM networks it is important to guarantee QoS. For this to be achieved, it is necessary to first setup a connection by reserving a certain amount of bandwidth and then providing the reserved bandwidth over multiple CDFs. To accomplish this task, the control phase must provide the following four basic functions:

- connection setup
- slot access
- slot confirmation
- connection release

Other functions such as connection re-adjustment, which enables a WT to change its original reservation request, can also be included.

4.4.2 Assumptions

The following assumptions are made for a office/home residential network scenario :

- There are only a limited number of simultaneous WTs.
- A WT cannot both transmit and receive data at the same time.
- Some information on the content of the ATM cell is available to the MAC layer, e.g. VPI, VCI and CLP.
- The unit of information transfer is assumed to be an ATM cell and an additional overhead corresponding to FEC/LLC etc. This unit is called an ATM slot.
- Only slow intra-cell movement is supported i.e. mobility is limited. No handover is allowed.
- Every WT within a cell can communicate with the BS serving that cell.
- A link between a transmitter and receiver can be assumed to be in one of the two states - "good" or "bad". Typically each state will be occupied for a significant amount of time after any transition.
- A known preamble must be sent by every transmitter to let the receiver equalizer adapt to the channel. This preamble is part of transmitter turn-around time.

4.4.3 Modes of operation for WTs

Before the details of various protocol functions are discussed, it would be useful to define the different modes of operation for a WT in a WATM network. The WTs can be classified into two categories as the following:

- Registered Wireless Terminals (RWTs): WTs who have registered themselves with the network.
- Non-registered Wireless Terminals (RWTs): WTs who have not registered themselves with the network.

RWTs can be further classified as :

- Non-networked Wireless Terminals (NNWTs) : RWTs who have registered into the network but have not had any service yet.
- Networked Wireless Terminals (NWTs) : RWTs who have been allocated bandwidth for at least one service.
- Active Networked Wireless Terminals(ANWTs) : NWTs who have been allocated bandwidth in the current CDF.
- Inactive Networked Wireless Terminals(INWTs) : NWTs who have not been allocated bandwidth in the current CDF.

Figure 4.4 shows the relationship between the different kind of users in the network.

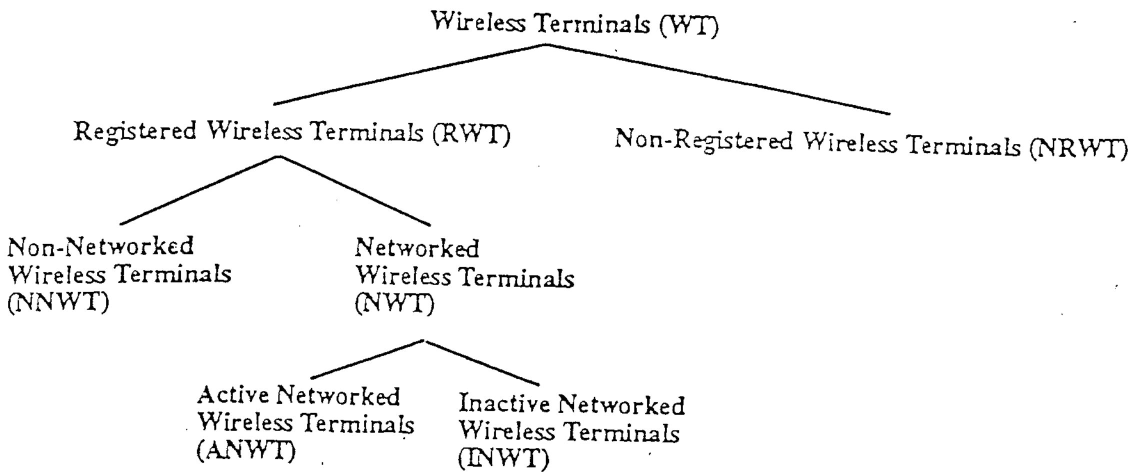

Figure 4.4: Modes of Operation of WTs

4.4.4 Salient functions of the MAC layer

Connection Setup Function

This function enables a WT to request a connection set-up with a certain nominal number of ATM slots. This is not really a function of the MAC layer, but is done by the Connection Admission Control mechanism in the parallel control plane.

The bandwidth allocation algorithm at the MAC takes the nominal bandwidth allocation as an input from the CAC and then allot slots on a per CDF basis. This is discussed in greater detail when the Slot Allocation policy is described.

Once the requested connection is granted, a nominal number of slots (bandwidth) is then allocated to the WT and this contract is honored for this service by allocating some ATM slots over every CDF over the duration of the service. To ensure that a WT does not get the entire bandwidth an upper bound is fixed on the number of slots that can be allocated to the WT.

Slot Access Function - Optimization of specific CDFs

Even though a nominal number of slots are reserved for every NWT in a CDF, there is a high possibility (due to the bursty nature of traffic) that in a particular CDF a service may not require the nominal bandwidth that has been allocated to it, giving rise to the possibility of a multiplexing gain. Hence, every NWT that wishes to transmit data in the next CDF must indicate this to the central controller (Base Station) the bandwidth desired in the next CDF. This is done by a polling method for INWTs and inband signaling for ANWTs. These are discussed in detail in later sections.

Slot Confirmation Function

In a slot access stage, there is a possibility that more resources are requested than are available i.e. the number of requested slots exceeds the total number of data slots in CDF (maximum CDF size). Hence, there is a need to explicitly indicate the number of slots allocated to each WT. This is provided by the slot confirmation function.

Connection Release Function

Once the allotted service has finished or doesn't require slots any more, a connection release function is used to free the reserved bandwidth.

4.5 Contention Free Multiple Access Scheme

A primary goal in the design of the MAC protocol was to keep the Data phase contention free. This is important to guarantee QoS. The control phase could be based on contention/contention free scheme, however since the number of users is expected to be limited, even the control phase was designed to be contention free. The following functionalities are thus required in the protocol:

- A base station which acts like a central controller and is entrusted with the responsibility of coordinating the access to the wireless channel.
- A reservation based multiple access scheme to allocate bandwidth to the wireless terminals as and when they request it.
- Signaling schemes from the active networked users to the base station to request bandwidth and from inactive networked users to base station to start transmissions. The scheme should be capable of conveying information regarding packets queued at each WT and the *priority* with which the packet needs to be served.
- Signaling schemes from the base station to the Networked users informing them of the allocated bandwidth and the slots allotted to them in the forthcoming CDF.

All of the above are described in detail in the following sections:

4.5.1 Superslot uplink signaling scheme for ANWTs

To implement the control channel for ANWTs, the piggybacking concept is used, i.e. signaling messages from the WTs to the BS regarding their BW requirements are sent along with the ATM data packets. Different schemes have been discussed in [19], including (a) the "inband" approach where signaling messages are carried in a separate ATM slot, (b) the "subslot" approach where signaling messages are carried separately in subslots and (c) the "piggybacked" approach where signaling messages are divided and piggybacked to the ATM slots. A "superslot" approach has been proposed recently by Gagan et al [20]. In this approach, two types of slots are used, namely superslots and normal ATM slots. A superslot is composed of a subslot attached to a normal ATM slot. Each ANWT at the start of its transmission cycle uses this superslot to transmit signaling information in the subslot portion and data in the rest of the slot. The succeeding slots for the same WT are normal ATM slots. Analysis and simulation has showed that this scheme gives much better performance and hence this scheme will be used here. Detailed analysis is given in Appendix A.

4.5.2 E_burst uplink signaling scheme for INWTs and NNWTs

To implement the control scheme for INWT and NNWT, a polling strategy is used to determine if a INWT or a NNWT has data to send and hence wants to start transmitting. In this scheme, the INWTs and NNWTs, when they wish to start transmitting , send a contention less short energy burst called "E_burst" in the "E_burst slot" allocated to them by the BS. The BS then assigns one ATM superslot to them, which can be used to transmit one ATM cell and signaling information stating the required number of slots in the next CDF. This strategy has minimal overhead and it can be shown that it causes a worst case delay of $2*CDF$ frame time for a INWT/NNWT who wishes to reenter the system. Also this scheme is quite simple to implement.

4.5.3 Downlink Signaling

The previous sections describe all uplink signaling from the WTs to BS. To ensure that all WTs have a common knowledge of the allocated slots in the each succeeding CDF the BS must broadcast downlink signaling information. The contents of the signaling information include the following:

- Number and position of data slots allocated in both downlink and uplink for each of the WTs
- The E_burst slots allocated to the INWTs and NNWTs
- Operations and Maintenance messages

All of the above discussion leads us to the following kind of structure for the MAC superframes:

4.6 MAC Frame Structure

In each MAC CDF there are four phases namely :

- BS_sig
- Down_data
- Up_data
- E_burst

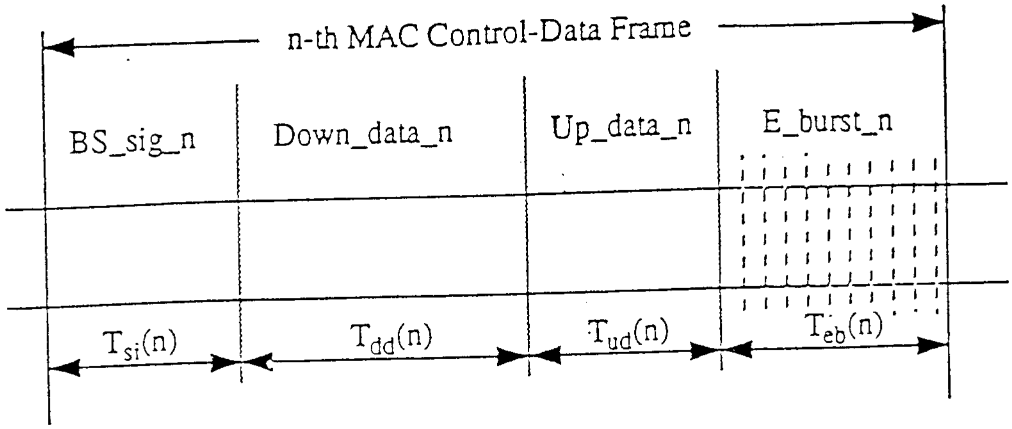

Figure 4.5: CDF structure

During the BS_sig phase the BS broadcasts all the downlink signaling information to the WTs. Following its signaling phase, the BS sends the downlink data in the Down_data phase. This minimizes the turnaround time of the BS. Next, the WTs transmit information in a prespecified order in the Up_data phase, which includes piggybacked signaling information using the superslot technique described earlier as well as the ATM data. Note that the size of the data phase (Down_data + Up_data) is variable and is limited to a maximum size, this kind of variable length CDF structure is optimal from delay considerations. The variable nature of the CDF reduces the delay suffered in low load situations and at the same time the limit on the maximum size of the CDF ensures that the maximum delay for a high priority service is limited. It can be shown that for a high priority service, the worst case delay is limited to $2 * CDF$ frame time.

During the E_burst phase, all WTs who did not send any data in the previous CDF, and were assigned specific E_burst slots by the BS, will transmit at their specific E_burst slot an energy signal if they wish to get any bandwidth in the next CDF.

As an example, suppose that an INWT wants to start transmitting again. It will first monitor the BS_sig phase of every CDF and see if there is an E_burst slot allocated to it by the BS. When it detects this E_burst slot, it will send an E_burst at the designated slot. After the BS receives the E_burst, it then allocates to the the INWT a superslot in the next CDF and notifies the WT of this in the BS_sig phase of the next CDF. The WT then sends its bandwidth requirements and one unit of data in this superslot.

The procedure for a connection setup for a NNWT is essentially the same. In the following each of these phases is described in more detail.

4.6.1 Phases of the MAC superframe

BS_sig phase

- Format : One or two ATM slots
- Purpose :
  - To specify the usage of slots in both Down_data and Up_data phases.
  - To specify the polling of INWTs and NNWTs during the E_burst phase.
  - To send operations and maintenance messages.
  - To send acks for other signaling messages.
- Description : The BS will announce the slot assignments for both the down_data and the up_data phases based on slot allocation done using a Slot Allocation Algorithm which takes into account both the reserved bandwidth for a service at the time of connection set-up (nominal slot

allocation) as well as the current needs of each service (dynamic state information) which is conveyed to the BS in each CDF by the WTs using the "Superslot" signaling. The slot allocation is described in detail in the next chapter. In this phase the BS essentially informs the WTs when to receive data in the downlink and when to send data in the uplink. During this phase the BS also specifies the polling sequence of INWTs and NNWTs during the E_burst phase in the current CDF. In addition it may also send out various kinds of signaling messages including operations and management messages (congestion indication etc) or acknowledgments for other signaling messages if any. Power saving is not possible during this phase because every RWT has to listen to the channel.

Down_data phase

- Format : A number of ATM slots
- Purpose : To deliver down-link ATM data to WTs in the order specified in the BS_sig phase.
- Description : The BS sends down-link data to the WTs in this phase and informs them about the same in the BS_sig phase described earlier. All NWTs, who do not have down-link data from the BS, can switch to power saving mode.

Up_data phase

- Format : A sequence of a superslot and zero or more normal ATM slots.

- Purpose : To enable NWTs to send their up-link signaling information and up-link data using superslots and normal ATM slots in the order specified in the BS_sig phase.

- Description : The ANWTs send signaling/data in this phase to the BS. All NWTs, who do not have assigned up-link data slot can switch to the power saving mode.

E_burst phase

- Format : A number of short energy bursts; each of a very small interval (E_burst time).

- Purpose : To enable INWT and NNWTs to indicate to the BS that they wish to regain an uplink channel.

- Description : The polling sequence of INWTs and NNWTs, who were allocated E_burst slots in the E_burst phase, is specified in the BS_sig phase of the current CDF, thus there is a one to one relationship between the between the position (time) of the E_burst and the WT which sent it. This type of polling has very low overhead if the number of INWTs and NNWTs at any time is limited, which is a fair assumption to make for the usage scenarios assumed for this kind of a system.

Chapter 5

Proposed Slot Allocation policy

5.1 Introduction

The increasing bandwidths in networks have made applications such as interactive voice and video communications feasible. Such services are characterized by high bandwidth requirements and real time constraints. Traditional communication applications such as file transfer and electronic mail are examples of "non real time" traffic. The performance metric for such applications is typically average delay and throughput. The characteristics of real time traffic, however, differ significantly. Most importantly, each real time traffic packet is characterized by a *deadline* by which time it must be received at its intended destination; a packet not received within this deadline is typically considered to be lost. Unlike traditional communication applications, many real time applications can tolerate such loss, provided that the amount of loss is kept within certain limits. In such cases, the primary performance metric for this class of traffic is the fraction of packets received within their deadline.

These different performance metrics suggest that protocols and architectures previously developed for traditional communication applications may not be well suited for real time traffic in wide area networks. Section 2.3 discusses the various service classes that need to be supported in a ATM network. We examine the use of different *link level scheduling policies* to support these classes of traffic and we also suggest a new scheme which is similar to some already existing schemes.

5.2 Review of different service classes and their requirements

The three main classes of traffic that are supported in ATM networks are the following:

- **Real Time Traffic**: This traffic class mainly supports voice and video (CBR and VBR) traffic and is characterized by a low cell transfer delay (CTD) and low cell delay variance (CDV) requirement. This class can however tolerate some loss. Real time packets are characterized by their deadlines. A real time packet which does not complete its transmission within its deadline is considered lost. The performance metric for real time traffic is thus the percentage of messages lost. Its also important that the jitter of these services be low.

- **Non Real Time Traffic**: This traffic class supports mostly data traffic and some non real time video traffic. This class is not delay or jitter sensitive, however it is sensitive to loss. The non real time traffic has no deadlines and hence can be buffered without loss; the performance metric for non real time traffic is average delay.

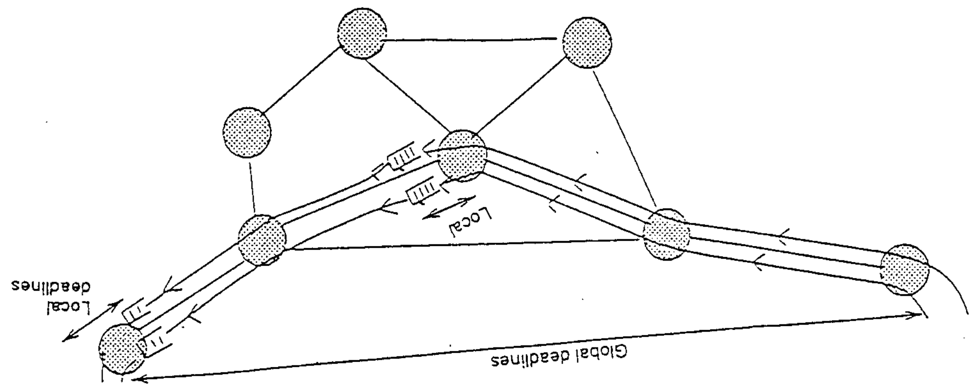

Figure 5.1: Global and local deadlines

5.3 Scheduling Issues for real time traffic

As mentioned earlier, real time traffic that is unable to meet a specified timing constraint is considered lost, regardless of whether it is eventually received at a destination node. Thus an end-to-end deadline can be specified along with a tolerable loss as a QoS metric for many types of real time multimedia traffic. One approach to meeting the end-to-end delay requirement is to distribute the end-to-end delay requirement across the various components of the communication path. Hence, the end-to-end or *global* deadline may be translated into *local deadlines* at intermediate network nodes as shown in Figure 5.1. By ensuring that local performance criteria are met, it is possible to guarantee that end-to-end delays are met for traffic delivered to the receiver. In the absence of local deadlines this guarantee is hard to provide because of variable nature of the network delays.

The danger with allocating the global deadline across intermediate nodes is that a particular packet that was unable to meet a local deadline may still have been able to meet the end-to-end delay constraint. This packet would nonetheless be dropped within the network. On the other hand, if no local performance criteria are enforced, a packet that would never make the end-to-end deadline may still continue to consume network resources unnecessarily. Thus, the particular allocation policy used must take both possibilities into account.

5.4 Different categories of the scheduling policies

Most of the scheduling policies can be classified into the following three broad categories:

- **First Come First Served(FCFS):** The FCFS scheduling discipline does not differentiate between real time and non real time traffic and all transmitted packets are sent on a first-to-arrive basis. This policy treats real time traffic poorly and the losses incurred are high [20]. Clearly this scheme is not suitable for a system where QoS guarantees need to be made.

- **Priority Based Schemes:** Under this scheduling discipline, real time traffic are given strict priority over non real time packets. Within the same priority class packets are transmitted in an FCFS manner. Packets of the same priority that arrive at the same time are transmitted in random order. There are many variants of this basic approach:

  - **EDD:** This scheme called the Earliest Due Date policy and was proposed by Jackson. Under EDD, each packet has associated with it a *due date*, typically interpreted as the time, by which to complete the transmission of the packet. At the time that the server becomes available, it is assigned the packet with the smallest (earliest) due date to transmit.

- STE/ML: This scheme called the Shortest Time to Extinction is same as the Minimum Laxity policy. These schemes are essentially same as the EDD scheme except that they never schedules tasks that are already past their due dates. Each arriving real time packet has a fixed real time deadline T_i. The laxity of a real time packet τ_i is defined to be the number of remaining slots until its deadline expires. A packets laxity always starts at τ_i and decreases by one at the end of each slot that it remains in the queue until either the packet is transmitted or its laxity reaches zero. In the latter case, the packet is removed without being transmitted. Panwar et al [21] have shown this scheme to be an optimal scheme for the discrete time G/D/1 class of queues, which is the commonly used model for statistical multiplexers in data communication systems.

The above priority based schemes are optimal for real time traffic for a single hop network, but the performance of non-real time traffic under these schemes is very poor and results in large delays. Also the above schemes are not suitable for multihop networks since under these schemes the end-to-end jitter tends to grow as a function of path length in networks, which makes these unsuitable.

To overcome this problem many variants of the STE scheme have been proposed, most notable among those are the following:

- FIFO+: This policy was proposed by Clark et al[22]. Each node estimates the average delay for all packets of a given class passing through it. Each packet carries a priority field that is initialized to zero as the packet enters the network. As a packet departs from the node, the difference between the queuing delay actually experienced at the node and the average delay for all packets of the class is added to the priority field. Thus, a positive value in the priority field indicates that the packet has spent more than the class average in the queues traversed so far. Conversely, a negative value indicates that the packet has received better than average service. At any queue, the packets which have exceeded their class average by the largest amount, are served first. This scheme has been shown to reduce jitter.

– **Hop Laxity (HL):** In this scheme the packets are served according to their hop laxity [23]. The hop laxity is defined as

$d_k = \frac{\tau_k}{M-k+1}$;

where d_k is the hop laxity at the k^{th} node in the path from source to destination, and M is the total number of hops between the source and destination. While waiting, d_k value of a packet reduces with a rate that depends on the number of hops left to traverse. At an output link, the packet with the lowest value of d_k is always transmitted next. Ties between packets with the same laxity are broken in favor of the earliest arrival. This local deadline based priority scheme captures the system time remaining until extinction, scaled by the number of hops that the packet still has to traverse. Thus, the packets tend to be "compensated" by a higher priority for above average delay suffered at earlier nodes. This scheme has been shown to be an optimal scheme for real time services in wide area networks from both the delay and jitter standpoint. However this scheme also is not suitable for non real time applications and gives very poor performance for such traffic.

- **J-EDD:** This scheme was proposed by Verma et al[24] and is called Jitter-EDD, and has the provable property that end to end jitter never exceeds that for a single node. This property results from the addition of an input traffic regulator which holds a packet until the time it is expected to arrive at the node. At that time it is released into the node to be scheduled.

- **Threshold based schemes:** These schemes are intended to give the designer explicit control over the performance tradeoffs between two traffic classes. These schemes essentially fall into two categories:

  - **Minimum Laxity Threshold (MLT):** Under MLT scheduling, priority is given to real time traffic when the minimum laxity among the queued real time packets falls below a certain threshold L ($0 \leq L \leq \tau$); otherwise priority is given to the non real time traffic. L may be chosen to achieve a particular tradeoff between the performance of two traffic classes. Again, packets within the same class are served in a FCFS manner and ties are broken randomly.

  - **Queue Length Threshold (QLT):** Under this policy, whenever the number of queued non real time customers exceeds the threshold value, T, non real time packets are given priority; otherwise priority is given to real time packets.

5.5 Slot Allocation Policy

A slot allocation (scheduling policy) policy as applicable to our system, is a control mechanism, which allocates a pre-defined number of slots among NWTs; each having multiple classes of traffic. In a BS oriented network, the policy is primarily implemented at the BS. It is important that the BS has detailed uniform information regarding the bandwidth requirements of all the WTs and the urgency with which they need to send data (important for real time services). However there exists a tradeoff between the performance and the overhead required to transmit and process this detailed information. In the following we consider what is essentially a modification of the HL policy discussed earlier. However in this policy there is a deadline associated with non real time traffic as well, which is kind of a timeout associated with each non real time packet. By varying this parameter we can improve the performance of the non real time traffic without sacrificing on the performance of real time traffic.

5.5.1 Problem Formulation

Conventionally, the slot allocation problem is formulated by considering only the uplink scheduling and ignoring the interdependency on downlink scheduling. This formulation is quite inaccurate especially when the size of the down-link phase is variable and dependent on the size of the uplink phase. In this case both the uplink and downlink scheduling need to be considered simultaneously and the problem can be formulated by considering two dependent schedulers : uplink scheduler and downlink scheduler.

The uplink scheduler considers $M(> 1)$ WTs sending packets to a BS over

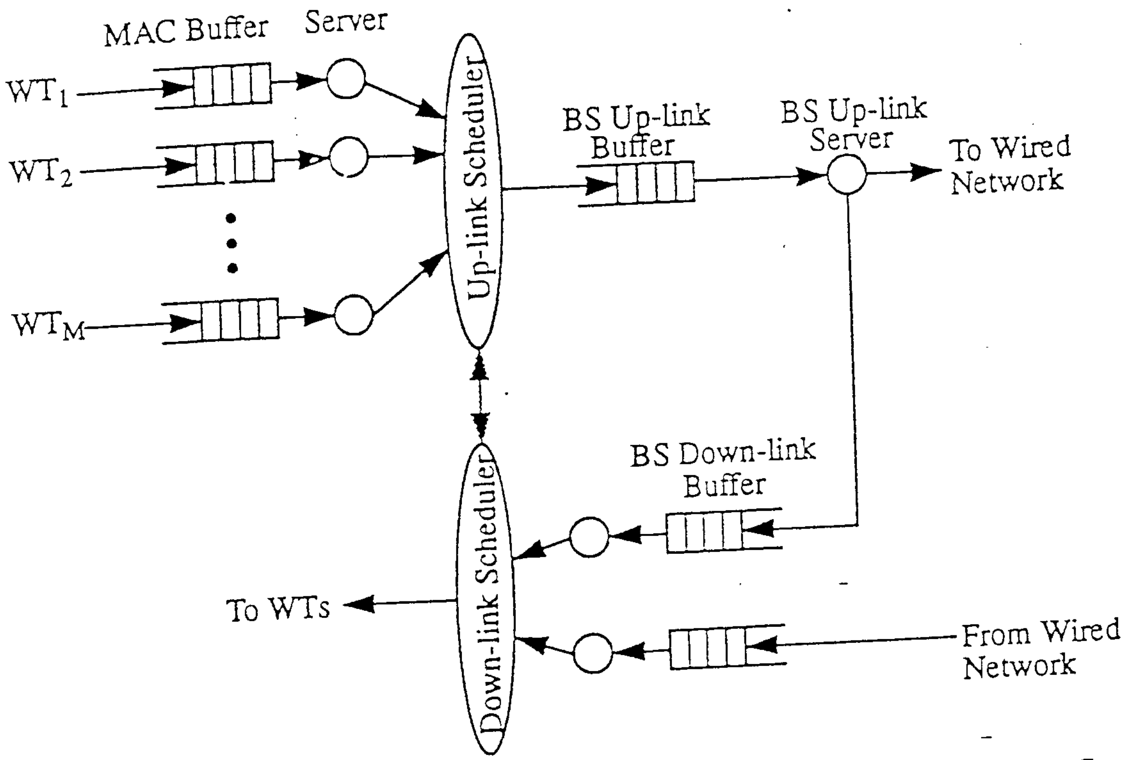

Figure 5.2: Uplink and Downlink schedulers

a time multiplexed channel during the Up_data phase of a CDF and the downlink scheduler considers the BS broadcasting packets to the WTs over the same channel during the Down_data phase. Without loss of generality the channel is assumed to be slotted, i.e. the channel time is divided into equal segments called slots. All messages consist of packets of equal length; the transmission time of a packet is one slot, and a packet transmission may only begin on a slot boundary. Only one WT or the BS is allowed to transmit during any particular slot and this is decided by the Slot Allocation policy at the BS and this information is broadcast to all WTs at the start of every CDF in the BS_sig phase.

Referring to the Figure 4.3, let $T_{dd}(n)$ and $T_{ud}(n)$ be the number of consecutive slots in the Down_data and the Up_data phase respectively of the n^{th} CDF. Prior to the beginning of these two phases, i.e in the BS_sig phase the BS informs each WT of the following :

- the number of consecutive packets and the beginning slot in which the WT is going to receive packets from BS during the Down_data phase.
- the number of consecutive packets and the beginning slot in which the WT is allowed to transmit during the Up_data phase. For the purpose of allocating slots in the n^{th} CDF, the BS is assumed to possess the following information prior to the start of the n^{th} CDF.
- the number of packets queued at each WT for uplink transmission for each priority class (if any) at the end of their data transmission in $(n-1)^{th}$ CDF, this information is sent by every ANWT in every CDF using superslot signaling described earlier.
- the nominal number of packets allowed at each WT.

- the number of packets queued at the BS for each priority class (if any) at the end of the $(n-1)^{th}$ CDF.

- the WTs that transmitted an energy burst during their E_burst slot in the $(n-1)^{th}$ CDF.

A policy for the BS is any function of the above mentioned information that allocates the down_data slots and up_data slots subject to the constraint $(T_{dd}(n) + T_{ud}(n) \leq T_{dmax}(n))$ i.e. the maximum frame size of CDF is limited. It is desired to determine policies that minimize the total mean delay and cell loss probability and maximize throughput while ensuring fairness among all WTs.

5.6 Proposed HL - based Slot Allocation Policy

The proposed slot allocation policy is based on the Hop-Laxity criterion (HL). Laxity is defined as the the Mean Residual Life Time (time before a cell is discarded) for a cell in terms of the number of ATM slots. Hop Laxity is the laxity per hop and is calculated as described in section 5.4. A mean residual life time can be allocated to each service (VCI) based on the delay sensitiveness of data. Even for a non-delay sensitive data a mean residual life time is assigned, although this is much higher than that for delay sensitive real time data.

The hop laxity calculation is done dynamically at each WT and the BS. A hop laxity vector can be generated based on the status of all the cells at the MAC queue in both the BS and the WTs. As described earlier it is important to consider along with laxity, the number of hops that the packet is expected to traverse before its final destination. The BS and the WTs maintain a look up table which associates with each source destination pair the number of hops that the packet is *expected* to travel. Estimation methods can be used to construct the above-mentioned lookup table. Here we just assume the existence of such a table and we are not concerned with methods on how to construct the same. The hop laxity vector can then be sent to the BS which allocates slots in a round robin fashion based on the hop laxity.

Upon allotment of data slots, cells are transmitted based on their hop laxity. Hence a cell with the lowest hop laxity is transmitted before other cells and so on.

5.6.1 Laxity to Priority mapping

To reduce the processing and bandwidth overhead during the BS_sig phase, the hop laxities at each WT and the BS are quantized and mapped on to a priority vector.

Priority is defined as the mean residual life time of a cell in terms of the maximum CDF size, hence *Priority = Laxity/Max. size of CDF (in slots)*

This priority mapping is of type (N_i, P_i) and is calculated every CDF at the WTs and BS, where (N_i, P_i) means N_i cells with priority P_i. A total of 16 priorities is used. The mapping is described below :

Uniform Quantizing with clipping : The relationship between the 4-bit priority level and the laxity is linear for all laxities less than Lclip where Lclip is the clipping laxity value beyond which all laxities belong to priority level 1111.

Non-Uniform Quantizer : In this scheme the relationship between the 4 bit priority level and the laxity is non-linear as shown below. Unlike the

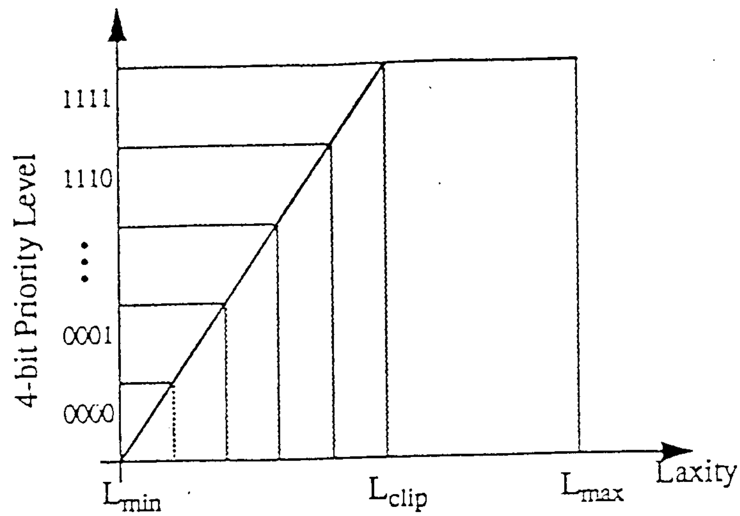

Figure 5.3: Uniform Quantizing

uniform case, the division of laxity is finer for small laxities. A common non-linear function used in the μ-law of digital PCM telephony is :

$y = sgn(x)\frac{ln(1+\mu|x|)}{ln(1+\mu)}$ where $-1 \leq x \leq 1; \mu = 100$

By normalizing laxity i.e. $x = \frac{(Laxity-L_{min})}{(L_{max}-L_{min})}$, we can use the above function for our laxity to priority transformation. Note that in our case, x is non negative and hence sgn(x) function in the equation can be omitted.

5.6.2 Uplink Signaling information

It is now possible to describe the uplink signaling information to be sent every CDF from the WT to the BS. For each priority level we sent the (N_i, P_i) mapping as described earlier. Here P_i is a 4 bit number and the number of priorities is 16 hence the total signaling overhead is 16*4 = 64 bits = 8 bytes. Hence the ATM superslot is as described in earlier sections.

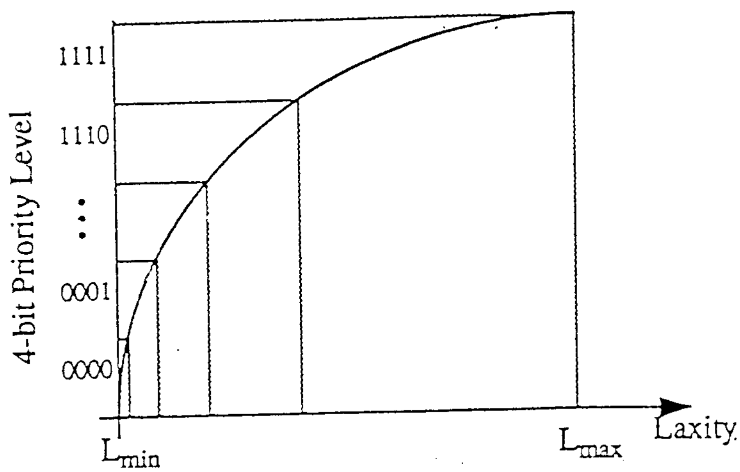

Figure 5.4: Non Uniform Quantizing

5.6.3 Slot Allocation algorithm

Thus the Slot Allocation Algorithm is as follows :

For $0 \leq i \leq M$ (0 corresponds to BS and 1 to M to the WTs), let priority_vector(i) be the priority vector sent to the scheduler by WT i. The details of the allocation policy at the scheduler can be illustrated as follows :

1. ATM Superslot assignment

   For $0 \leq i \leq M$

   If $(E\_burst(i) > 0)$*, then*

   grant 1 ATM superslot to node i;

2. Nominal Slot assignment

   For $0 \leq i \leq M$

   If $(Total\_packets(i) \leq Nominal\_packets(i))$*, then* *grant Total_packets(i) to node(i);*

clear priority_vector(i);

else

grant Nominal_packets(i) to node(i);

subtract Nominal_packets(i) from priority_vector(i) sequentially

starting from the entry corresponding to P_{min}*;*

3. Slot Assignment according to Priority

For all non-zero vector(i), $0 \leq i \leq M$*,*

assign the rest of slots based on priority in a round-robin fashion

until no more spare slots are left ;

Nominal BW allocation is assumed to be done by the CAC layer in the wireless control plane (Figure 3.3) and the allocation is given to the above algorithm.

5.7 E_burst Allocation

A E_burst slot is allocated to each inactive networked WT and to some NNWTs in each CDF.

5.8 Buffering Strategy

The cells are stored on a FIFO priority queue - cells with same priority are stored in a FIFO discipline, but cells with higher priority are ahead of those with lower priorities. The prioritisation is based on the HL criterion described earlier.

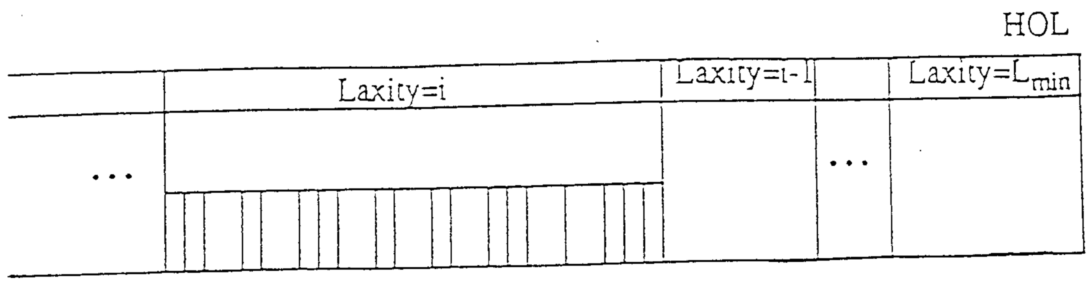

Figure 5.5: Buffering Scheme at the MAC layer

This calculation is done prior to the start of the transmission period of each WT in every CDF.

5.8.1 Buffer Overflow Control

The buffer overflow control strategy is based on the ABR congestion control scheme. Each ABR source has a current rate called the Actual Cell Rate (ACR) that is bounded by the Peak Cell Rate (PCR) and the Minimum Cell Rate (MCR). In our system when congestion occurs i.e. the queue length at a WT/BS exceeds a certain threshold L_{MAX}, the rate of the ABR sources is dropped to the MCR (which can be zero). When the network comes out of congestion i.e. the queue length at a WT/BS falls below a threshold L_{MIN}, the rate for the corresponding ABR source is restored back to the ACR. This scheme is a simplified version of the ER (explicit rate) scheme commonly used in ATM networks, whenever congestion occurs in the system, ER is set to the MCR. The thresholds L_{MAX} and L_{MIN} are simulation parameters.

Chapter 6

Simulation, Results and Conclusions

6.1 Simulation Description

In the discrete event simulation we simulate the performance of a cell within the system, which consists of a Base Station (BS) and five wireless terminals (WTs). The number of active networked users is thus five. The channel is of broadcast type and can introduce packet errors based on simulation parameters. Since strong forward error correction (FEC) is being used, only packet errors are assumed. Separate links have been simulated for uplink and downlink traffic. It is assumed that the nominal bandwidth allocation is done by a CAC algorithm in the control plane and this is known to the Slot Allocation Algorithm at the Base Station. Equal load to all WTs is assumed. Traffic sources described later are used to generate traffic at each WT. New connection set up requests are not allowed. Thus the simulation setup captures a snap shot of typical MAC layer operation. The TDMA frame structure is implemented using a set of timers which trigger the events corresponding to each phase in the Control Data Frame. The communication between the WTs and the BS is simulated using shared memory between various entities.

6.1.1 Base Station

The BS module consists of the following main modules which do the following tasks:

- Slot Allocation: The Slot Allocation module (SA) allocates slots for both the uplink and downlink data channels in every CDF based on a slot allocation algorithm described earlier. It reads the *priority_matrix* into which each WT and the BS write information about the number of cells queued and their respective priorities.

- E_burst Allocation: The E_burst allocation module allocates E_burst slots to all the inactive networked wireless terminals, by writing into a shared memory, which is read during the E_burst period by all the INWTs.

- The queue module at the base station receives and buffers packets from the WTs and transmits them in the downlink slots. The buffering strategy used at the BS is the same as the one used at each WT and is described earlier. The laxity and priority calculation modules calculate the laxity and priorities of the cells queued at the base station and write into the priority_matrix at the start of the downlink data phase.

- The BS module implements the CDF frame structure using a set of timer modules which trigger events corresponding to each phase in the CDF namely the BS_signaling, Data and the E_burst phase. These durations are calculated in every CDF by the Slot Allocation module.

- The BS module also send a congestion indication in the BS_signaling phase if the buffer length at the BS exceeds the threshold L_{MAX} described earlier.

There are five Wireless Terminals (WTs). Each WT module consists of three main modules :

- A traffic sources which generates CBR, VBR and ABR traffic as described later in this chapter. The loads and the traffic mix i.e. the composition of the traffic is a simulation parameter. Also the time to expiry associated with each service is a design parameter.

- A transmit module which queues the generated packets in a buffer and transmits them in the slots allocated to them by the BS. The buffering and the congestion control strategy as described earlier are implemented in this module.

- A physical link which is capable of introducing packet errors. The packet error probability is a simulation parameter.

6.2 Traffic Sources

As mentioned earlier, ATM networks must support various communications services, such as data, voice, and video and each having different traffic characteristics. To evaluate the performance of such networks, accurate source modeling is required. The purpose of this section is to describe the traffic models used for this simulation for different service classes.

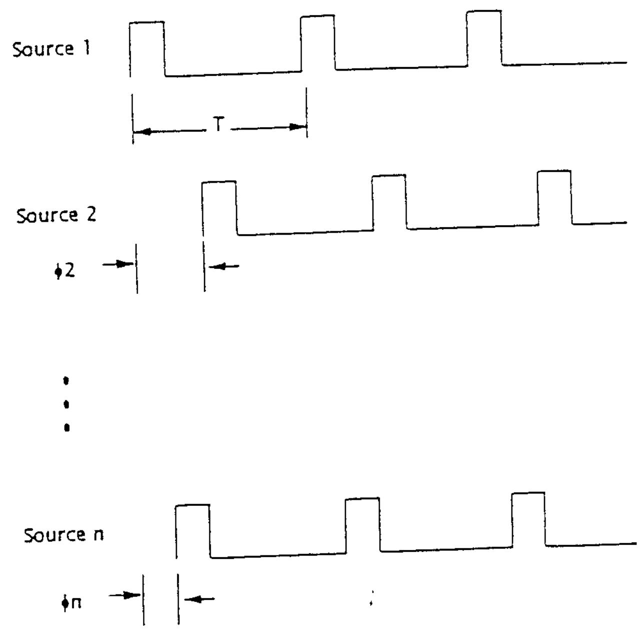

Figure 6.1: Randomly Phased CBR Sources

6.2.1 CBR Sources

These sources generate traffic at a constant rate derived from the mean cell rate. This model is implemented using a traffic generator which generates pulses at a constant rate which can be specified by the user. These pulses trigger the generation of data structures representing the MAC packets. Each mobile can have up to a maximum of three such sources. The three sources are identical (can have different parameters as well) but they start the transmission at some random phase in the interval (0,T) as shown in Figure 6.1.

6.2.2 VBR Sources

These models are applicable to video scenes with relative uniform activity levels such as video-telephone scenes showing a person talking. These are based on an Auto Regressive Moving average model (ARMA) [26]. In this model a frame is generated every 1/30th of a second and the cell rate during the n^{th} frame is given by the following:

$Y(n) = 0.8781 \times Y(n-1) + 0.213 \times w(n) \times M$, where

$w(n)$ is a Gaussian random variable with a mean of 0.572 and variance of 1.0, and M is the mean cell rate per frame of the video source. These sources are simulated using a traffic source that generates pulses at a constant rate at the start of each frame and in the n^{th} frame the number of cells generated is given by $Y(n)$. There can be a maximum of three such sources at each WT. The first frame occurrence of each of the sources is randomized over the interval of a frame. Once initialized, the sources keep their individual frame synchronizations and the cells are generated at a rate according to the above-mentioned formula.

6.2.3 ABR sources

The ABR sources can also be modeled as a superposition of identical ON-OFF mini sources [3] (two state Markov modulated Poisson processes). In the ON state the sources generates packets at a uniform interval. The number of packets is geometrically distributed with a mean equal to 18 ATM cells (mean Ethernet packet length) and the ON periods are Poisson distributed.

6.3 Simulation Parameters

Bit Rate: This is the raw bit rate at which data is transmitted at the physical layer.

T_S: OFDM Symbol Time = 4 bytes of information corresponding to 16 OFDM carriers, each carrying 2 bits of information and the guard time.

T_R: Rise time of the radio transmitter.

T_F: Fall time of the radio transmitter.

T_E: The time required to send an E_burst over the channel.

T_T: Transmitter turnaround time - time associated with synchronization and sending training bits etc.

FEC: The number of bytes used for the FEC overhead attached to the MAC payload.

ns: The number of bytes used for sending the signaling message.

Max Load: This is the load for which the load on the system is considered to be 100% and is calculated in packets per second. It is assumed here that all data transmission is between WTs and the BS and further that approximately half the slot in each frame are allotted for the uplink transmissions.

Load: This is the load on the system at any point of time and is expressed as a fraction of the Max Load.

Stop Time: The time at which the traffic sources stop generating traffic.

Mean CBR cell rate: This is the mean rate at which CBR traffic is being generated by each service at each WT.

Number of CBR sources: This is the number of CBR sources that are generating traffic at each of the networked users. This can vary between zero and three.

Res Life Time - CBR: This is the residual time (time to expiry) at the start of service for each CBR packet being generated.

Mean VBR cell rate: This is the mean rate at which VBR traffic is being generated by each service at each WT.

Number of VBR sources: This is the number of VBR sources that are generating traffic at each of the networked users. This can vary between zero and three.

Res Life Time - VBR: This is the residual time (time to expiry) at the start of service for each VBR packet being generated.

Inter Frame Time: This is the duration between successive VBR frames.

Inter Cell Delay - ABR: This is the time between successive ABR cells during an ON period.

Mean Delay between bursts - ABR: This is the mean time for an OFF period for an ABR source.

Number of ABR sources: This is the number of ABR sources that are generating traffic at each of the networked users. This can vary between zero and three.

Res Life Time - ABR: This is the residual time (time to expiry) at the start of service for each ABR packet being generated.

ATM Superslot length: The time required to transmit an ATM superslot.

ATM slot length: The time required to transmit an ATM slot.

BS_sig slot length: The time required to transmit the BS signaling information.

Buffer Size - MT: The buffer size used at each mobile.

Buffer Size - BS: The buffer size used at the BS.

Packet error prob: The probability that an ATM packet will have to be discarded due to errors.

Number of Registered Users: The number of users who have registered with the network.

Number of Networked users: The number of users who have been granted at least one service.

Number of Data Slots: The maximum number of slots that can be allocated during one CDF.

6.4 Results

6.4.1 Performance Metrics

- Delays: As can be seen from Figure 6.2 , the delay performance of the protocol is very good. At low loads the mean delays are $\leq$ 0.5ms. Even at peak load, the mean delay for CBR service is $\sim$ 6ms, which is adequate to meet the QoS guarantees. Also the performance of delay insensitive (ABR) service is very good, which is one our primary design goals.

- Jitter: It can be seen from Figure 6.3 that the standard deviation of delays for CBR packets is limited to less than $\sim$ 2ms at high load and at low load is considerably lower.

- Cell Loss: From Figure 6.4 we can see that the cell loss ratios are zero for low to moderate loads. At very high loads also the cell loss ratio is only $\sim 8 \times 10^{-5}$. Also note that this is a dramatic improvement over the Cell Loss when the ABR congestion control scheme is not implemented.

- Deadline Expiry: An important QoS metric for real time services is the number of packets that had to be discarded, since their deadlines had expired. The simulations show that in our protocol, even at very high loads, this is zero. Thus our original design goal of minimizing both the average delay for delay insensitive service as well as the packet loss in real time services due to deadline expiry has been met successfully.

6.4.2 Design Parameters

- Maximum CDF Size: From Figure 6.3 , we can see that mean delays decrease as the maximum number of data slots per CDF increase. This means it is better from mean delay point of view to have larger CDF sizes, however as indicated earlier the worst case delay suffered by an "urgent" service is 2*CDF size and this grows as the frame size increases. Thus there are two contradicting requirements. Hence in our view a suitable size for the maximum number of data slots per CDF = 150. The CDF size at high loads in this case is $\sim$ 2ms, which is similar to many existing TDMA based protocols (e.g. PACS).

- Buffer Sizes: As can be seen from the Figure 6.5 the mean queue length $\leq$ 15, even at very high loads. Also the instantaneous queue lengths at peak load (Figure 6.5) never increases beyond $\sim$ 85. Thus at each mobile the buffer size of $\sim$ 100 seems adequate.

6.5 Conclusions and Future Work

We have demonstrated here a reservation based MAC protocol capable of supporting a diverse traffic mix. The key to effective and efficient bandwidth allocation is the availability of detailed state information for each packet queued at the WTs and the BS to the Central Controller (in this case the base station itself), as is done in the proposed protocol. The bandwidth allocation using this information results in significant statistical multiplexing gain. The transmission of this information from the mobiles to the BS is done in a bandwidth efficient manner using a "superslot" signaling scheme. The performance of the protocol both from delay and cell loss performance metrics is very satisfactory, as seen from the simulation results. We have also been able to extract certain key design parameters such as the maximum CDF size and the buffer sizes at each mobile.

However the protocol doesn't yet have provisions for mobility and hand-off support. So future work needs to be done in this area. We haven't looked at the control and management aspects of the MAC layer in the control plane, notable among these signaling and Connection Admission Control, which need to be studied. It is desirable to have a more sophisticated congestion control scheme as well. Also we are currently assuming a fixed amount of FEC for each packet, which entails a large overhead. Thus it is desirable to have a more sophisticated adaptive FEC/ARQ scheme. The amount of FEC bytes used should be adjusted dynamically depending on the state of the channel. Also from the simulation point of view we need to model the physical layer more accurately.

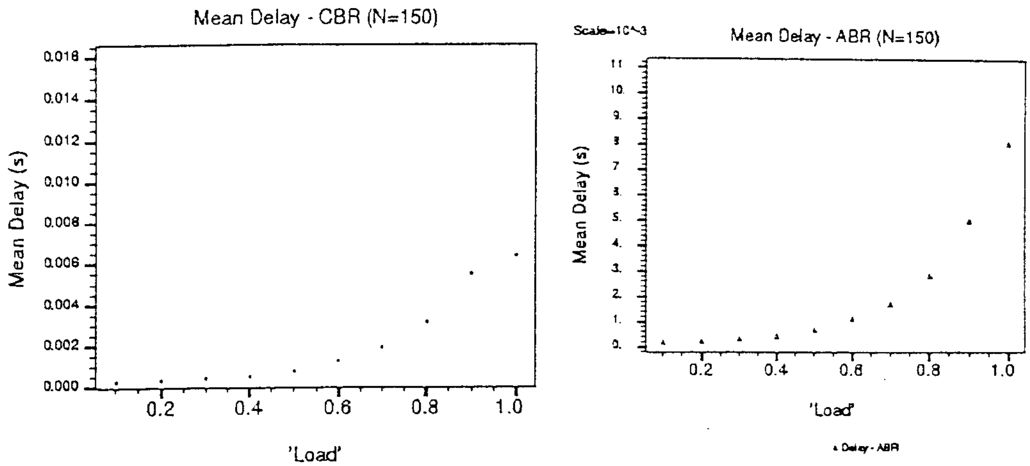

Figure 6.2: Mean Delays - CBR and ABR

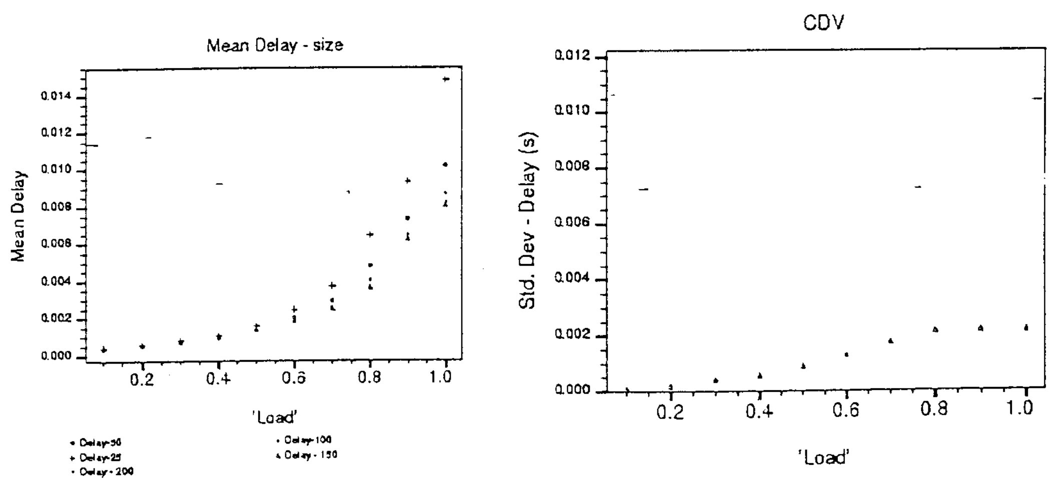

Figure 6.3: Mean Delays - Size and CDV

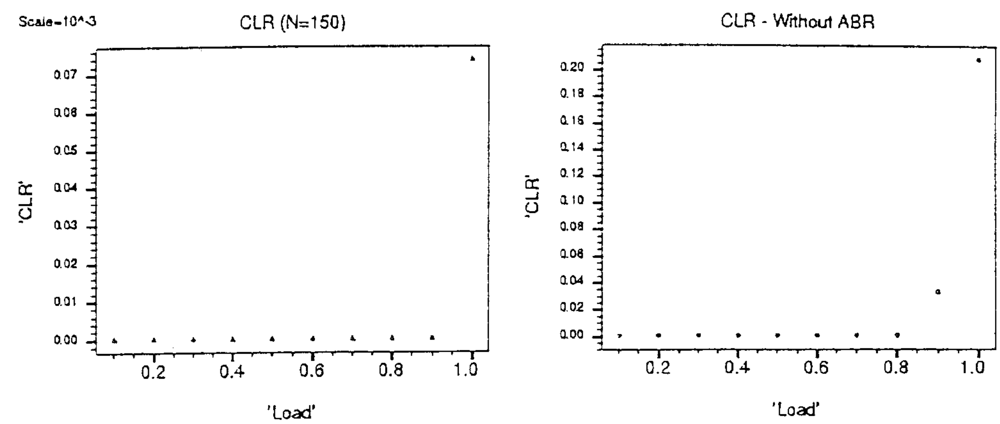

Figure 6.4: CLR Comparison

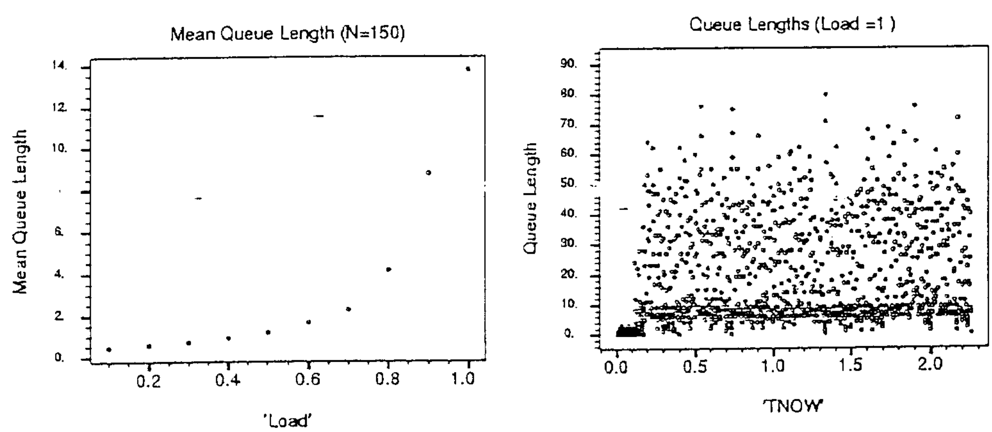

Figure 6.5: Mean and Instantaneous Queue Lengths

Appendix A

Overhead comparison of uplink signaling schemes

A.1 Introduction

The overhead and useful bandwidth comparison is done here for three existing signaling schemes [19] used by the WTs to convey their dynamic bandwidth requirements in every frame to the base station. Also a new scheme called the superslot scheme [20], which gives a significantly better performance in terms of overhead and useful bandwidth is presented.

The analysis also considers the overhead due to Forward Error Correction (FEC). The FEC overhead is assumed to be 20 bytes. It is also assumed that all schemes use the same amount of FEC overhead. For the subslot strategy the use of 20 bytes FEC overhead is somewhat questionable and inefficient. However in the interest of limiting the analysis, a uniform assumption is made.

It is also assumed that all the signaling information is attached to only one ATM cell. Furthermore the overhead used for signaling is of a fixed size (ns) bytes and is used as a parameter in the analysis (ns is varied from 2 to 16 bytes). The rationale for these assumptions is based on the observation that after a WT announces its intention to transmit during the E_burst interval, it gets assigned one slot in the next CDF. During that slot, it should be able to specify all its signaling information. Finally, this analysis includes the E_burst interval for calculation of overhead. This has necessitated the consideration of two parameters described earlier the total number of networked users (N_{NU}) and the total number of active networked users (N_{ANU}).

A.2 Definitions

T_S: OFDM Symbol Time = 4 bytes of information corresponding to 16 OFDM carriers, each carrying 2 bits of information and the guard time.

T_R: Rise time of the radio transmitter.

T_F: Fall time of the radio transmitter.

T_E: The time required to send an E_burst over the channel.

T_T: Transmitter turnaround time - time associated with synchronization and sending training bits etc.

N: The number of ATM slots in a CDF.

N_{NU}: The number of networked users.

N_{ANU}: The number of active networked users.

ns: The number of bytes used for sending the signaling message.

A.3 Reservation Schemes

A.3.1 Inband

In this scheme reservation messages for each wireless terminal wishing to transmit data are carried in a separate ATM slot. The size of each slot is equal to 54 bytes (53 byte ATM cell + 1 byte MAC header). A 20 byte FEC is attached to it at the physical layer. The size of the CDF is then given by:

$$L_I = (N_{NU} - N_{ANU}) \times T_E + (N_{ANU} + 1) \times T_T + N \times 74 \times T_S/4$$

The reservation message overhead to a first approximation is dependent on the number of WTs, N_{ANU} because every WT requires one slot to transmit the reservation message:

$$O_I = \frac{(N_{ANU} \times 74 \times T_S/4)}{L_I}$$

The overhead including E_burst is given by :

$$OE_I = O_I + \frac{(N_{NU} - N_{ANU}) \times T_E}{L_I}$$

The useful bandwidth (the bandwidth used to send the MAC payload) is given by:

$$BW_I = \frac{(N - 1 - N_{ANU}) \times 53 \times T_S/4}{L_I}$$

A.3.2 Piggyback

In this scheme reservation messages for each wireless terminal wishing to transmit data, are carried piggybacked to the MAC payload. The size of a MAC PDU is considered equal to $(54 + ns)$ bytes (this allocation of ns bytes is made because of the earlier assumption that all signaling information is sent in the first slot). A 20 byte FEC is attached to each slot at the physical layer.

The length of the CDF for this case is:

$$L_P = (N_{NU} - N_{ANU}) \times T_E + (N_{ANU} + 1) \times T_T + N \times (74 + ns) \times T_S/4$$

The reservation overhead is independent to a first order of the number of sending WTs and is given by:

$$O_P = \frac{N \times ns \times T_S/4}{L_P}$$

The overhead thus increases linearly with the number of bytes used for sending reservation messages(ns).

The overhead including E_burst is given by:

$$OE_P = O_P + \frac{(N_{NU} - N_{ANU}) \times T_E}{L_P}$$

The useful bandwidth for this case is given by:

$$BW_P = \frac{(N-1) \times 53 \times T_S/4}{L_P}$$

A.3.3 Subslot

In this scheme MAC reservation messages are transmitted separately in subslots. The size of a MAC PDU is equal to 54 bytes. The size of a subslot is considered equal to ns bytes for reservation messages. The number of subslots required is equal to the number of sending WTs, N_{ANU}. To each subslot we need to attach a 1 byte MAC header and a 20 byte FEC which results in a significant overhead.

The length of the CDF is dependent on the number of subslots and is:

$$L_S = (N_{NU} - N_{ANU}) \times T_E + (N_{ANU} + 1) \times T_T + (N_{ANU} \times (ns + 21) + N \times 74) \times T_S/4$$

The reservation overhead increases linearly with the number of active WTs,

$$O_S = \frac{N_{ANU} \times (ns+21) \times T_S/4}{L_S}$$

The overhead including E_burst is given by:

$$OE_S = O_S + \frac{(N_{NU} - N_{ANU}) \times T_E}{L_S}$$

The useful bandwidth for this scheme is given by :

$$BW_S = \frac{(N-1) \times 53 \times T_S/4}{L_S}$$

A.3.4 Superslot

A new scheme was proposed by Gagan et al [20] and is called the superslot scheme. In this scheme there are two kinds of slots - superslots and normal

ATM slots. Superslots comprise of a subslot(ns bytes) attached to a normal ATM slot. Each transmitting WT at the start of its transmission period uses this superslot to transmit a reservation message in the first ns bytes and data in the rest of the slot. The succeeding slots for the same WT are normal ATM slots. WTs who have indicated that they want to regain the channel, by transmitting during their E_burst period are allocated one ATM superslot during the subsequent frame and they can get more slots in subsequent CDFs by sending signaling information for their bandwidth requirements in the subsequent frames. The number of superslots is thus equal the number of active WTs, N_{ANU}. The 20 bytes of FEC instead of being attached to each slot is now attached to the superslot, thus saving on overhead.

$$L_{SP} = (N_{NU} - N_{ANU}) \times T_E + (N_{ANU} + 1) \times T_T + (N_{ANU} \times ns + N \times 74) \times T_S/4$$

The overhead for this scheme is proportional to the number of sending WTs, N_{ANU} but does not include the extra 20 bytes due to FEC and is given by:

$$O_{SP} = \frac{N_{ANU} \times ns \times T_S/4}{L_{SP}}$$

The overhead including E_burst is given by :

$$O_{ESP} = O_{SP} + \frac{(N_{NU} - N_{ANU}) \times T_E}{L_{SP}}$$

The useful bandwidth is given by:

$$BW_{SP} = \frac{(N-1) \times 53 \times T_S/4}{L_{SP}}$$

$$O_S = \frac{N_{ANU} \times (ns+21) \times T_S/4}{L_S}$$

A.4 Simulations and Results

Extensive simulations in MATLAB were done. Here is a summary of the results:

- The overhead of the Inband scheme O_I increases very rapidly with the number of transmitting WTs, N_{ANU}.
- The overhead of the piggybacked scheme O_P is almost constant with respect to N_{ANU} but increases linearly with the size of the signaling message (ns bytes) and is quite large for large values of ns (e.g.16 bytes).
- The overhead of the subslot scheme O_S increases with the number of active WTs, N_{ANU}, but the increase is not as rapid as in the Inband case. This scheme essentially suffers due to the FEC overhead of 20 bytes which needs to be attached to every subslot.
- The overhead O_{SP} for the superslot scheme is significantly lower than all of the above schemes. The overhead due to the superslot scheme for $N = 151$ and $ns = 8$ bytes is less than 0.5%, which is really low. Also even when the number of the WTs goes up in the cell the overhead is just about 0.8%, thus the scheme scales up pretty well.

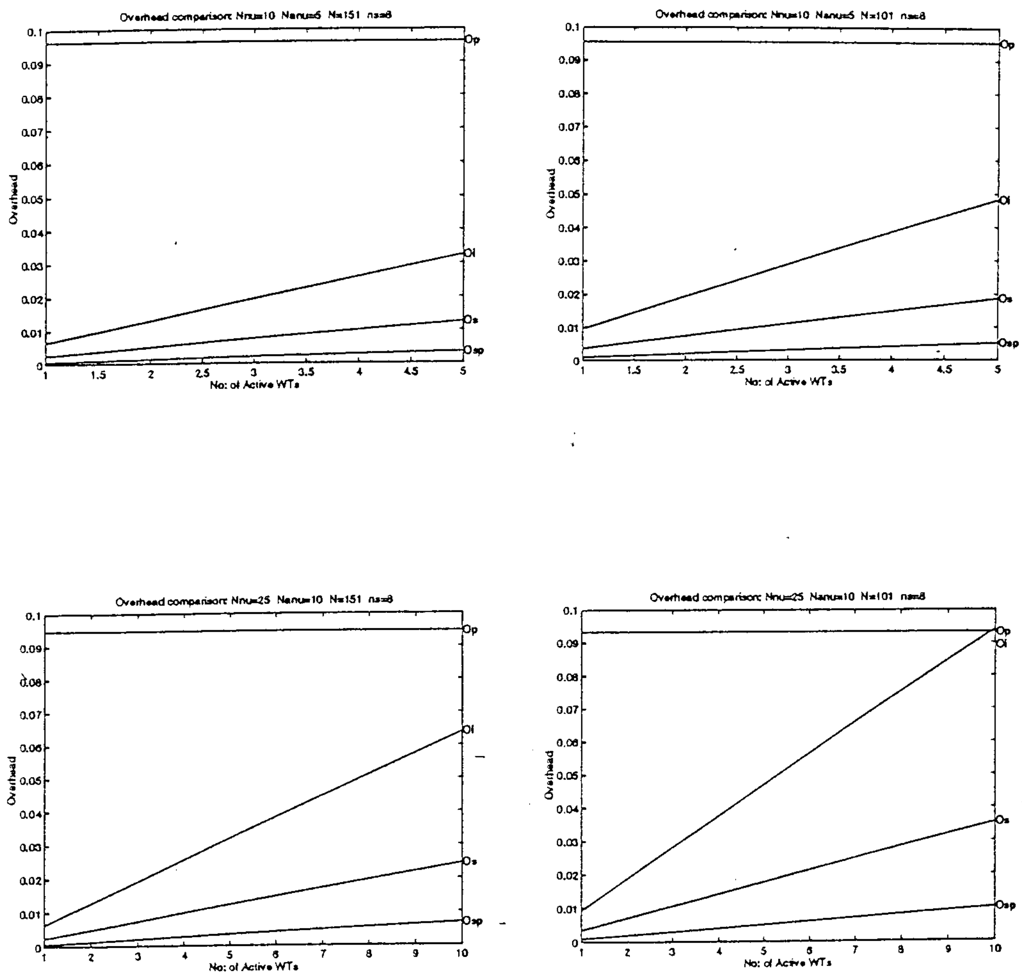

Figure A.1: Overhead Comparison

Bibliography

[1] D. Raychaudhari and N. Wilson, "ATM Based Transport Architecture for Multiservices Wireless Personal Communication Network", *IEEE JSAC*, vol. 12, pp. 1401-1414, Oct. 1992.

[2] J. Y. LeBoudec, "The asynchronous transfer mode: a tutorial", *Comp. Networks and ISDN Sys.*, vol. 24, pp. 279-309, 1992.

[3] M. DePrycker, *Asynchronous Transfer Mode - Solution for Broadband ISDN.* Prentice Hall, 3rd. ed., 1995

[4] R. Jain, S. Kalyanaraman and R. Viswanathan, "The OSU Scheme for Congestion Avoidance in ATM networks Using Explicit Rate Indication", Proceedings WATM'95 First Workshop on ATM Traffic Management, Paris, December 1995

[5] D. Raychaudhari, "Wireless ATM Networks: Architecture, System Design and Prototyping", *IEEE Personal Communications Magazine*, pp. 42-49, Aug. 1996.

[6] H. Hansen, " Connection Management Functions of a Private Wireless ATM Network", *Master's Thesis*, Helsinki University of Technology, Mar. 1996.

[7] Y. Du, "Wireless ATM Networking Using Distributed Switch Architecture", Patent Application, Dec. 1993.

[8] G. Water and J. Kruys, "Overview of High Performance Radio LANs", presented ETSI RES 10, Jan. 1996.

[9] V. O. Li, "Performance Evaluation of Multiple-Access Networks : Introduction and Issue Overview", *IEEE JSAC*, Vol. 5, No. 6, July 1987.

[10] E. Ayanoglu, K. Y. Eng and M. J. Karol, "Wireless ATM: Limits, Challenges, and Proposals", *IEEE Personal Communications Magazine*, pp. 18-34, Aug. 1996.

[11] J. Porter and A. Hopper, "An Overview of the ORL Wireless ATM System", *Proc. IEEE ATM Workshop, Washington. DC.*, Sept. 30 - Oct. 1, 1995.

[12] K. C. Chen, "Medium Access Control of Wireless LANs for Mobile Computing", pp. 50-63, *IEEE Network*, vol. 8, 1994.

[13] H. Chou, C. Lee and K. C. Chen, "Group Random Addressed Polling with reservation for Wireless Integrated Service Network", pp. 618-622, *Proc. of 6th IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'95)*, Toronto, Canada, 1995.

[14] A. Acampora, "Wireless ATM: A Perspective on Issues and Prospects", *IEEE Personal Communications Magazine*, pp. 8-17, Aug. 1996.

[15] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard and B. Ramamurthi, "Packet reservation multiple access for local wireless communications", *IEEE Trans. Commun. COM - 37*, pp. 885-890, 1989.

[16] M. J. Karol, Z. Liu and K. Y. Eng, "An Efficient Demand-assignment Multiple Access Protocol for Wireless Packet (ATM) Networks", *Wireless Networks 1*, pp. 267-279, 1995.

[17] M. Umehira et al, "An ATM Wireless Access System for Tetherless Multimedia services", *Proc. ICUPC'95*, Tokyo, Japan, Nov. 1995.

[18] C. Y. Ngo and S. Hulayalkar, "Wireless ATM LAN - Deliverable D3.1", version 1.0, Aug. 1996.

[19] A. Hettich, "Calculation on reservation message schemes", RWTH-TN, Jun. 1996.

[20] G. Singh and S. N. Hulayalkar, "Bandwidth and overhead comparison among various reservation schemes", PRB-TN, July 1996.

[21] R. Chipalkatti, J. F. Kurose and D. Towsley, "Scheduling Policies for Real-Time and Non-Real-Time Traffic in a Statistical Multiplexer", *Proc. IEEE Infocom'89,* (Ottawa, Canada), pp. 774-783, 1989.

[22] S. Panwar, D. Towsley and J. Wolf, "Optimal Scheduling Policies for a Class of Queues with Customer Deadlines to the Beginning of Service", *J. of the ACM*, Vol.35, pp. 832-844, Oct. 1988.

[23] D. D. Clark,S. Shenker and L. Zhang, "Supporting real-time applications in an integrated services packet network: architecture and mechanisms", *J. of ACM*, pp. 14-26, Aug. 1992.

[24] H. Schulzrinne, J. Kurose and D. Towsley, " An Evaluation of Scheduling Mechanisms for providing Best-Effort, Real-Time Communication in Wide-Area Networks", *Infocom 1994*, Toronto, Canada, 1994.

[25] D. Verma, H. Zhang and D. Ferrari, "Delay jitter control for real-time communications in a packet switching network", *IEEE Tricom '91, April 1991.*

[26] B. Maglaris, D. Anastassiou et al, "Performance Models of Statistical Multiplexing in Packet Video Communications", *IEEE Trans. on Communications*, Vol. 36, No. 7, pp. 834-844, July 1988.