ENHANCEMENTS OF CFDAMA PROTOCOL FOR MULTI-SERVICE APPLICATIONS IN A GEO BROADBAND SATELLITE NETWORK

By

OLUMIDE B. IBIKUNLE

B.Sc., Obafemi Awolowo University, Nigeria, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

ELECTRICAL AND COMPUTER ENGINEERING

THE UNIVERSITY OF BRITISH COLUMBIA

AUGUST 2005

Abstract

The bursty behavior and the diverse quality of service (QoS) requirements of the component traffic generated at the user terminals (UTs) in current broadband satellite networks (BSN's) have been shown in numerous research projects to necessitate the use of dynamic medium access control (MAC) protocols such as the combined free/demand assigned multiple access (CFDAMA) protocol for efficient statistical multiplexing of the traffic from the several UTs on the uplink channel. Several predictive algorithms have been proposed to enhance the CFDAMA protocol performance by estimating the dynamic trends in the bursty UT traffic and allocating the uplink capacity based on those trends. In this thesis, we provide enhancements to the predictive algorithms of the CFDAMA protocol using a two-frame traffic momentum estimation strategy in a time division multiple access (TDMA)/multi-frequency (MF)-TDMA uplink frame format. Moreover, the performances of the CFDAMA protocol variants were enhanced for an application environment consisting of a mixture of real-time (RT) and non-real-time (NRT) applications. For RT applications we take advantage of packet-level application profile information obtainable from the real-time protocol (RTP) and real-time control protocol (RTCP). The resulting protocol relies on the cross-layer application-profile description provided to implement intelligent resource request signaling (RRS), dynamic capacity allocation (DCA), and a priority-based local capacity allocation scheme. Simulation results for the proposed enhancements, when compared with previous variants of CFDAMA, display significant performance improvements in terms of minimizing the delay and delay jitter or distributing these performance metrics between the application types in the UT to meet the different QoS requirements.

Contents

List of Tables

List of Figures

List of Acronyms and Abbreviations

ABR	Available Bit Rate
AF	Assured Forwarding
APID	Application Identification
APINFO	Application Information
ATM	Asynchronous Transfer Mode
BSN	Broadband Satellite Network
BSS	Broadband Satellite System
BSA	Broadband Satellite Access
BWA	Broadband Wireless Access
CAC	Class Application Count
CBR	Constant Bit Rate
CDMA	Code-Division Multiple Access
CFDAMA	Combined Free/Demand Assigned Multiple Access
CID	Class Identification
CL	Controlled Load
DA	Demand Assignment
DAMA	Demand Assigned Multiple Access
DCA	Dynamic Capacity Allocation
DiffServ	Differentiated Service
DSL	Digital Subscriber Line
DSPP	Doubly Stochastic Poisson Process
DVB-RCS	Digital Video Broadcasting- Return Channel System
EF	Expedited Forwarding
EPR-CFDAMA	Enhanced Predictive Combined Free/Demand Assigned Multiple Access
FA	Fixed Assignment
FAMA	Fixed Assignment Multiple Access
FBN	Fractional Brownian Motion
FDMA	Frequency Division Multiple Access
FGN	Fractional Gaussian Noise

F-TDMA	Fixed-Time Division Multiple Access
FTP	File Transfer Protocol
GEO	Geo-Synchronous Earth Orbit
GES	Gateway Earth Station
GOP	Group of Pictures
GS	Guaranteed Service
HAS	High Source Aggregation
HTTP	Hyper-text Transfer Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
ISP	Internet Service Provider
ISL	Inter-Satellite Link
LAN	Local Area Network
LCAS	Local Capacity Assignment Scheme
LEO	Low Earth Orbit
LRD	Long Range Dependence
MAC	Medium Access Control
MCS	Master Control Station
MEO	Medium Earth Orbit
MF-TDMA	Multi-frequency Time Division Multiple Access
MPEG-2	Motion Picture Expert Group 2
MPLS	Multi-Protocol Label Switching
MSA	Medium Source Aggregation
NCS	Network Control Station
NRT	Non-Real Time
ON-OFF	On-Off
Opt-CFDAMA	Optimized Combined Free/Demand Assigned Multiple Access
PDF	Probability distribution Function
PDU	Protocol Data Unit
PES	Personal Earth Station
PHB	Per-Hop Behavior

PRNRTT	Partly Real Time/ Non Real Time Traffic
PR-CFDAMA	Predictive Combined Free/Demand Assigned Multiple Access
QoS	Quality of Service
QPSK	Quadrature Phase Shift Keying
RA	Random Assignment
RF	Radio Frequency
RR-CFDAMA	Round-Robin Combined/Free Demand Assigned Multiple Access
RRS	Resource Request Signaling
RSVP	Resource Reservation Protocol
RT	Real Time
RTCP	Real Time Control Protocol
RTP	Real Time Protocol
RTT	Round Trip Time
SFRP	Superposed Fractal Renewal Process
SI	Service Information
SNRTT	Strictly Non-Real Time Traffic
SOHO	Small-Office/Home-Office
SRD	Short Range Dependence
SRTT	Strictly Real Time Traffic
TCP	Transmission Control Protocol
TDMA	Time Division Multiple Access
TWTA	Traveling Wave Tube Amplifier
UDP	User Datagram Protocol
UT	User Terminal
VBR	Variable Bit Rate
VHSA	Very High Source Aggregation
WAN	Wide Area Network
W-CFDAMA	Weighted Combined Free/Demand Assigned Multiple Access

Acknowledgment

My deepest gratitude goes to God, the source of all I have and to my supervisor, Dr. V.C.M Leung for his support, advice and supervision during the course of the research work leading to this thesis. I am also immensely grateful to Dr. Zhifeng Jiang, for helping in kick-starting the project work and for his unwavering attention all through the research. I should also express my thanks to all members of the examination committee for reviewing my thesis.

This research was supported in part by a grant from the Canadian Institute for Telecommunications Research under the NCE program of the Canadian Government, a grant from ComDev International, by the Canadian Natural Sciences and Engineering Research Council and the Canadian Space Agency under grant CSAPJ 223232-98, and by the OPNET University Program.

I would also like to thank all those who live in my heart; my mother Florence Aduke, my father, Albert Ibikunle and my brothers Dotun, Gbemi, Seye and and sister, Banke. I owe them many thanks for their faith in my dreams. Finally, I should acknowledge my late half-sister, Sade Sadiku for sharing some of her life with me; I dedicate this thesis to her memory.

Chapter 1 Introduction

Broadband satellite systems (BSS's) are becoming an integral and essential part of the global communications infrastructure. The exponential growth of the Internet, resulting in an increased demand for communications services by geographically remote users and the continued convergence of computing and communications, together with a rapid growth in the development of end-user applications such as interactive multimedia applications have all created unique communications demands that can be effectively met with the use of networking systems based on broadband satellite communications.

Satellite systems have the long-known advantage of wide coverage area, high-data rate and multipoint-to-multipoint communications facilitated by their broadcast capability. All of these unique features can be harnessed to meet the essential requirements of most of the new generation of users. Recent advancements in transmission technology have also led to the availability of low-cost satellite earth terminals with the potential effect of increasing the applications of broadband systems in meeting personal communications demands. While many BSS's currently deployed use ATM-like switching with onboard processing systems to provide full two-way services to and from earth stations or user terminals, several existing BSS's still employ non-regenerative on-board technologies.

1.1 The basic features of a broadband satellite system

The features of typical broadband satellite network (BSN) architecture are not significantly different from those of any typical satellite system. The broadband characteristics of the applications that are supported in the system, frequencies of operation of the systems and the data-rates achievable in the broadband satellite channels uniquely differentiate the broadband systems from any typical satellite system. A BSS can be divided into two inter-working segments: the space segment and the ground segment.

1.1.1 The space segment

The space segment of a BSS usually consists of a single satellite or a constellation of satellites in the geo-synchronous earth orbit (GEO) or low earth orbit (LEO). Inter-satellite links (ISL) in the constellation enable global coverage for the network especially in LEO constellations where the coverage area of a single satellite is limited.

The configuration of the satellite payload is dependent on the kinds of protocols or services that are to be supported in the system as well as the available technology for payload systems design. BSS's with on-board processing capabilities usually have a suite of signal processing subsystems to regenerate the original baseband digital signal and perform a number of signal processing functions such as multiplexing/demultiplexing, channel encoding/decoding and modulation/demodulation on the signal before retransmitting the signal to the earth stations. They also possess a host of advanced antenna systems features such as multi-beam antenna systems in addition to on-board switching capabilities to optimize power-to-coverage area of the system. The switching capability permits satellite adaptability to terrestrial ATM systems or networks [1], [2], [3].

1.1.2 The ground segment

The ground segment in a BSS is made up earth stations which could either be user terminals (UT's), gateway earth stations (GES's) that may be further connected to other legacy public and/or private networks and the network control station (NCS) or the master controller station (MCS). The GES acts as the interface between the satellite system and the terrestrial systems that are used to extend the service area of the system.

The gateway interface units present in the systems provide external network connectivity allowing seamless inter-working capabilities with terrestrial networking technologies. They perform the translation between the BSN's internal protocols and the standard protocols of the terrestrial world [1]. They must support several protocols standards such as ATM user network interface (ATM-UNI), frame relay UNI (FR-UNI), narrow-band integrated digital network (N-ISDN) as well the TCP/IP protocol suite.

The number and placement of the GES's in both GEO and LEO systems depend on traffic demand, performance requirements and other international regulatory issues. The UT's are the end terminals in the satellite network. The UT interface units (U-TIU) in UT's support several protocol standards adapting the application traffic in them to the satellite network. It includes a host of physical layer functionalities such as channel coding, modulation and demodulation, multiple access signaling and resource reservations and radio frequency (RF) functions.

The UT's offer different data transmission rates depending on the kinds of traffic sources that are supported. The UT's in a BSS can be grouped into the following three classes based on on the range of services they provide and the QoS values achievable [4]:

i. Fixed UT's providing a full range of multimedia services with a high QoS factor;

ii. Nomadic UT's with a similar service spectrum as fixed UT's but with a lower QoS factor due to power limitations of the antenna;

iii. Mobile UT's, offering a limited range of services that are characterized by a lower QoS factor than that of the other two groups.

The control actions in a BSS are performed by NCS/MCS. The NCS/MCS performs resource allocation functions together with configuration management, performance management and traffic management functions for the satellite media. The number and location of the NCS/MCS also depend on the size of the network, required area of coverage and other international regulatory issues. BSSs are usually deployed in two modes which are:

i. As a broadband satellite access (BSA) network;

ii. As a broadband satellite access/core network.

Figures 1.1 and 1.2 show the two basic configurations of BSS's.

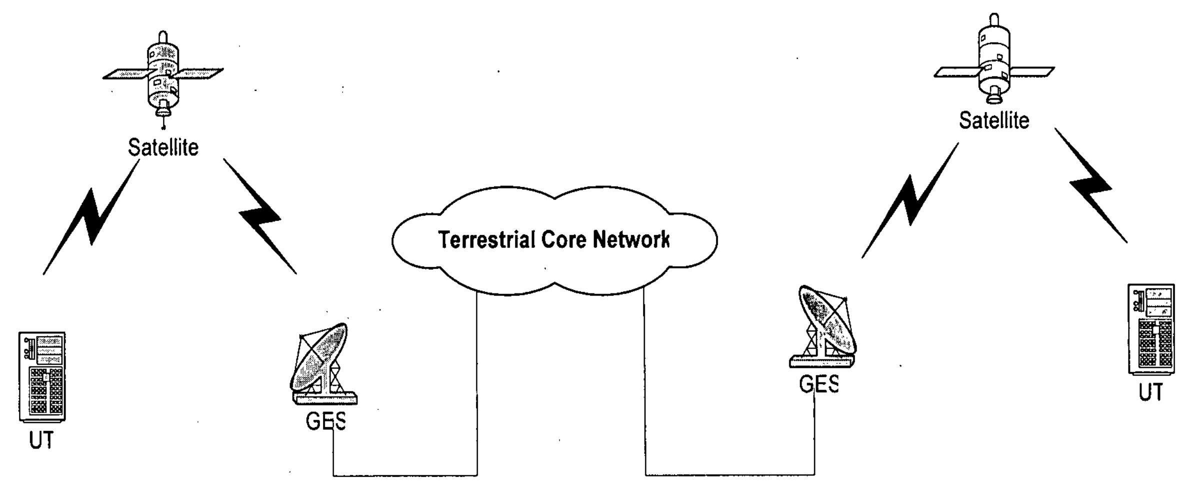

Figure 1.1: BSS deployed as a BSA network

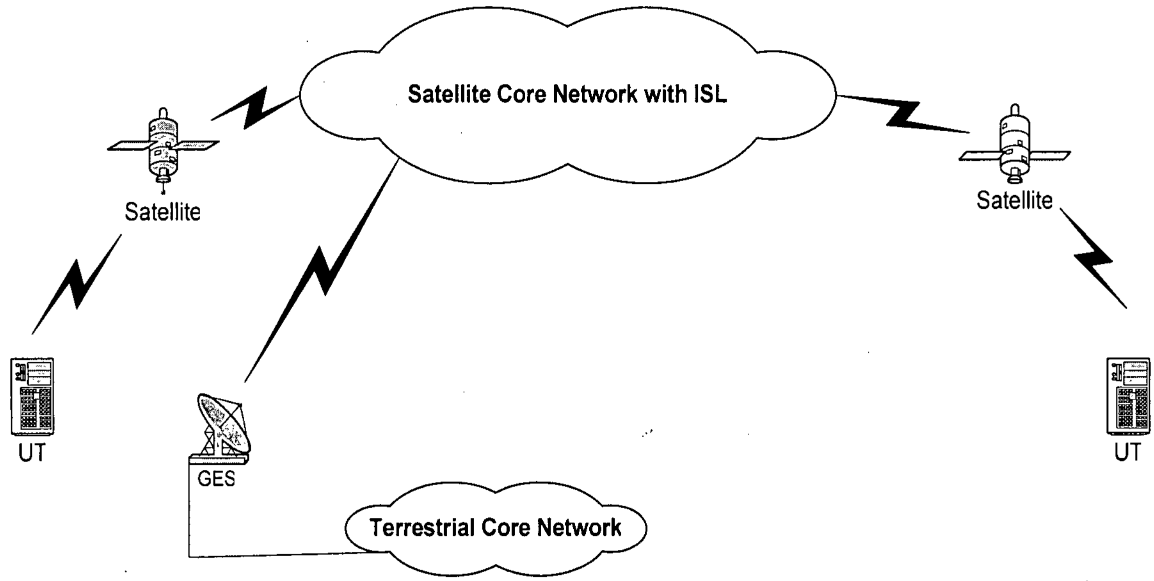

Figure 1.2: BSS deployed as access and core network

In the BSA network, the traffic signal transmitted by a UT is received by the satellite on the return link and retransmitted to a GES. The transmission of the original signal to an intended destination UT or to a nearby vicinity proceeds via a GES to a terrestrial network (which acts as the core network) on a forward link. Some flexibility can be permitted in the satellite coverage by means of a multi-beam architecture on the satellite antenna system. Thus, several areas served by a single downlink beam may be served by multiple uplink beams or vice versa. The flexibility achievable comes at the expense of an increased complexity of the satellite payload system due to the necessity for beam switching and some other onboard processing functions that may be performed on the signal, all of which imposes limitations on the power or service life of the satellite. In the access/core network, the signal sent from a UT or other sources and received by the satellite is transmitted via inter-satellite links (ISL) through the satellite network (which often provide onboard processing and switching) to the satellite serving the recipient UT or to the terrestrial core network via a GES.

BSA systems are becoming a necessary approach in improving telecommunications access especially for geographically remote customer areas where there is no terrestrial infrastructure. Until recently, the deployment of BSA systems for bidirectional communications service for broadband applications such as interactive multimedia services was hindered by the lack of economical satellite-based return links to transport user traffic [3]. However, the recent development of return-channel technologies and standards for satellite-based multimedia communications has effectively removed this obstacle [3],[10]. As a result, BSA systems can now compete with other wireless access technologies such as broadband wireless access (BWA) and digital subscriber line (xDSL) in terms of customer size and profile. The ranges of applications for which BSA systems are becoming increasingly popular include the following [3]:

i. Broadband connectivity for personal/enterprise communications e.g. small office/home office (SOHO) applications;

ii. Coverage area extensions for residential customers or backbone infrastructure applications for service operators;

iii. Wide-area telecommunications services such as telecommuting, tele-education etc.

In all of these application situations, two-way BSA systems can provide cost-effective broadband service solutions.

1.2 Application/transmission protocols and transmission platforms in a BSS

IP-based multimedia applications are becoming increasingly universal owing to the increasing popularity of the Internet for commercial, recreational, political and educational purposes. Thus, it is essential for BSA systems to interoperate seamlessly with the terrestrial IP networks and to be compatible with IP-based technologies and protocols. Also, because of the diversity of the universally available local area network (LAN) technologies (e.g. IEE 802.3 and ATM-LAN) and wide area network (WAN) technologies (e.g. frame relay, ATM etc), it is equally essential for the BSA systems to be able to inter-operate with terrestrial networks based on non IP-based technologies and protocols. For BSA systems to compete successfully with these terrestrial broadband systems, especially for IP-based applications, several functionalities must be provided to ensure adequate QoS guarantees that will complement the traditional best-effort service of IP. Additionally, BSA systems must also achieve efficient capacity utilization. This often means that BSA systems must provide and utilize mechanisms to intelligently manage the traffic offered by UT's while efficiently utilizing the network resources.

1.2.1 Application/transmission protocols

Several BSA systems support multimedia and other application systems with the following protocols and digital platforms;

i. Transmission control protocol/Internet protocol (TCP/IP);

ii. User datagram protocol/Internet protocol (UDP/IP);

iii. Real-time protocol/Real-time control protocol(RTP/RTCP);

iv. ATM protocol;

v. DVB-S digital platform.

The TCP/IP protocol suite was originally designed for use by Internet applications on terrestrial links to provide end-to-end reliable data delivery. It however has significant throughput performance deficiencies over satellite channels as a result of certain characteristic features of satellite links such as high packet loss rates and long round-trip times (RTTs). Higher packet loss rates on satellite channels lead to TCP triggering its network congestion control mechanisms to reduce its congestion window. Also, the long RTT characteristic of satellite links results in a slow congestion window increase by TCP's during the slow-start phase. Several link-layer and end-to-end solutions have been proposed to combat these deficiencies of TCP in a satellite system [3],[5],[6],[7],[8],[9].

The RTP/RTCP suite typically runs on the UDP/IP and is one of the most popular protocols for real-time application transport. The RTP represents a new style of protocol following the principles of application level framing and integrated layer processing proposed by Clark and Tennenhouse [11]. This integration enables RTP to be malleable to provide the information required by applications.

Several BSA systems make use of the ATM protocol as the basic transmission protocol. The ATM protocol acts on the principle that a virtual channel should be set-up between two points whenever such a communication need arises. The ATM protocol describes a transmission format in which data is formatted into fixed size (53 bytes long) packets called cells. These features of the ATM protocol contrast to the TCP/IP protocol in which messages are transmitted in packet form and each packet may reach the recipient via a different route. The ATM protocol enables data transmission through various media. It also defines several application service categories which are fundamental to resource allocations and traffic handling procedures. The service categories defined are constant bit rate (CBR), variable bit rate (VBR) (real-time VBR (rt-VBR), non-real-time VBR (nrt-VBR)), available bit rate (ABR) and unspecified bit rate (UBR).These service categories are associated with different QoS requirements [3], [12].

1.2.2 Transmission platforms

The digital video broadcasting-satellite (DVB-S) platform

DVB-S is a scalable platform that determines new standards for digital satellite broadcasting and for the supply of multimedia services. In general, the protocol includes the following [2]:

i. Methods for source coding of voice and video signals (typically MPEG-2/MPEG-4) as well as data signals;

ii. The method of including additional information for the appropriate configuration of the decoder and its synchronization (service information (SI)) [2] [10];

iii. Protection of the signal by the Reed-Solomon outer code, by interleaving and by the inner code (in the form of a punctured convolutional code).

The basic unit for the data transmission using the DVB-S platform is a 188-byte container. The system can be made flexible and adaptable to a variety of environments (depending on the power, antenna size, or error rate) by varying the efficiency of the inner code. Transmission is performed via QPSK modulation. DVB-S uses the MPEG-2 standard for the source sound and video signals thus reducing the rate of signal transmission. The typical transmission rates depend on the information content of the video signal and on the quality demands. Table 1.1 below shows the quality of transmission values and MPEG-2 bit rates [2].

Table 1.1: Transmission rate for MPEG-2 encoded applications in a DVB-S

Source data rate after MPEG-2 Encoding(Mb/s)	Application
6	Broadcast quality
4	Most users detect no visible degradation
2-3	Sports
1.5	VHS quality for film material

DVB-S also provides an open standard (DVB-return channel satellite (DVB-RCS)) for the return channel in BSA systems for interactive multimedia applications [10]. Several other satellite return-channel techniques have been developed or are currently being developed to provide an efficient and economical distribution of broadband multimedia services via BSA systems [1], [2].

1.3 Multiple access techniques in broadband satellite systems

Multiple access control is a fundamental function in satellite systems. The transmission capacity in the return link is often shared by numerous UT's in the delivery of the UT traffic to the recipient UT's or the appropriate GES. Hence, the basic function of multiple access techniques is to control access to the return channel in order to achieve certain pre-defined system's performance goals. These goals include efficient capacity utilization and the provision of appropriate link-layer service guarantees to individual UT in the BSN. Several multiple access protocols have been devised and evaluated in the literature [12]-[16]. Generally, multiple access techniques can be classified into three categories. These are:

i. Fixed assignment multiple access (FAMA) techniques;
ii. Demand assignment multiple access (DAMA) techniques;
iii. Random access (RA) techniques.

In the FAMA technique, the satellite resource in the form of return link access time or frequency bandwidth are permanently assigned to individual UT's and GES's each of which transmits data from its terminal queue for a pre-assigned duration or at a pre-defined carrier frequency. FAMA techniques have the advantage of simplicity but lack flexibility and reconfigurability. Thus, because the traffic patterns of the terminals are usually dynamically changing, FAMA schemes usually result in capacity wastage and are most effective for satellite networks composed of a small number of terminals with stable and predictable traffic patterns.

DAMA techniques utilize the dynamic characteristics of the traffic generated at the UT's to dynamically allocate the return channel bandwidth resource and are thus best suited for terminals with bursty traffic characteristics. In many variations of the DAMA technique, unused portions of the bandwidth resource are typically re-assigned to UT's with additional resource demands thereby ensuring efficient utilization of the bandwidth resource. A mechanism is usually provided for the UT's to request service and for the assignment to be appropriately calculated. This can represent an overhead of capacity and complexity. DAMA schemes can be achieved using a NCS/MCS as the central controller or the access scheme can be implemented in a distributed arrangement in which the capacity allocation and control functions are performed by the geographically distributed UT's in the BSN. In several new generation satellite systems with onboard processing systems, DAMA schemes can be implemented as an onboard function.

RA techniques are statistical access techniques in which access to the return channel capacity by subscribers' UT's is based upon bursty transmission of packets of data without any specialized control scheme. RA techniques are suitable for satellite networks with UT's having low average traffic rates. Packet collisions in the channel result in retransmissions. Random techniques are often used as a part of the control/reservation protocol in DAMA techniques [12], [13], [14].

Multiple access transmission formats are employed on the return link of a satellite in conjunction with the demand assignment or fixed assignment multiple access techniques to implement the channel access control and/or resource sharing protocol. The commonest transmission formats that are employed on a satellite channel are:

i. Frequency-division multiple access (FDMA);
ii. Time-division multiple access (TDMA);
iii. Code-division multiple access (CDMA).

In the FDMA transmission format, the satellite's return link capacity (in the form of bandwidth) is divided into sub-bands, each of which represents the bandwidth allotted each terminal in the network. The partitioning of the channel bandwidth into sub-bands can either be carried out statically, as in the FAMA multiple access techniques or it can be done dynamically as in the DAMA techniques.

A disadvantage of FDMA (especially in the DAMA technique) is the problem of inter-carrier interference (caused by inter-modulation effects, especially in earlier transponder designs employing traveling wave tube amplifiers) which occurs as a result of the non-linearity in the characteristics of most transponders employed in satellite payloads. This often necessitates the introduction of guard bands between the allotted sub-bands or operating the amplifiers on the satellite payload at signal levels much less than the capacity of the transponder thereby limiting the system's throughput.

In the TDMA transmission format, the transmission duration is divided into time frames and each frame is further divided into time slots. Each time slot represents the smallest transmission time duration and can be fixed to be the equal to the transmission duration of a single packet in packet satellite systems. In the FAMA form of the TDMA (termed fixed TDMA (F-TDMA)), each terminal is statically allotted time slots in each transmission frame and each terminal is synchronized to know when to transmit. Guard times are usually allowed between time slots to prevent collisions on the channel.

In the DAMA form of the TDMA transmission format, each terminal makes reservations for the slots available in each frame according to the amount of traffic generated locally at the terminal. The uplink transmission frame is divided into two sub-frames which are:

i. The control sub-frame;

ii. The data sub-frame.

The control sub-frame is used primarily for signaling purposes such as time-slots reservations, assignment information disseminations and other control functions. The boundary between these two sub-frames can be made movable to allow for better capacity utilization in situations where some of the slots in the control sub-frames are not utilized or to reduce contention for the control slots when there are free slots in the data sub-frame [15]. The reservation scheme can either be explicit or implicit. In explicit reservation, a single reservation slot is assigned to each active station in every TDMA frame. In implicit reservation, stations compete for the reservation slots using a number of techniques many of which are random in nature, e.g., slotted Aloha [15].

Actual data transmissions occur in the data sub-frame. The TDMA format generally offers a higher system throughput performance than the FDMA format owing to the absence of inter-modulation effects allowing for better transponder power utilization. In recent systems, the transmission formats employed is a hybrid of the FDMA and the TDMA formats, a transmission format referred to as the multi-frequency TDMA or MF-TDMA. Here the return channel bandwidth resource is a frequency-time map of slots and allocations to terminals consisting of sequences of frequency-time slots. MF-TDMA achieves higher transponder power utilization characteristics. A major disadvantage of the classical TDMA scheme is the need for sufficient peak power at the UT's to support the high instantaneous channel transmission rate. This requirement can only be met using large terminal antennas. MF-TDMA enables transmission at lower data rates thus enabling smaller and low-power terminals to transmit over the channel while providing the advantages of the basic TDMA transmission format.

1.4 Motivation and objectives

Despite significant disadvantages of satellite systems such as long delays, high bit-error rates and limited bandwidth/power resources, BSN's are emerging as a leading force in the delivery of adequate quality broadband services to customers in geographically diverse or remote areas. IP-based services such as IP-based multimedia applications require diverse QoS requirements because of the different characteristics of the data traffic they generate. While several advanced satellite systems based on the ATM protocol have a variety of techniques to adequately characterize the various traffic types in multimedia applications (for effective QoS provisioning), IP-based systems need additional protocols to provide and manage the service interface definitions for effective resource management.

In BSA systems, the medium access control (MAC) protocols employed in managing the resource allocations is essentially a QoS provisioning mechanism. The bursty nature of many multimedia/Internet applications often requires implementing predictive MAC protocol algorithms that can efficiently distribute the bandwidth resource while the varied characteristics of the component traffic in many broadband applications necessitates embedding methods in the MAC protocols for an intelligent association of application characteristics with appropriate QoS and resource management procedures.

The combined-free demand assignment multiple access (CFDAMA) protocol was first proposed as a MAC protocol to provide improved capacity utilization and end-to-end delay and jitter for a broad range of applications in a BSS. Several qualitative evaluations in numerous research studies have yielded results that shows how the CFDAMA can be employed in implementing some aspects of the differentiated service (DiffServ) framework by employing different techniques in the re-allocation of the free capacity that often exists in the return links of BSN's supporting bursty Internet-type applications in the UT's [15],[16]. An efficient predictive CFDAMA protocol scheme proposed in [15] estimates the first and second-order changes in burstiness observed in the self-similar UT traffic to dynamically allocate the free capacity in the CFDAMA scheme. The resulting simulations results showed significant performance improvements over previously existing CFDAMA protocol algorithms [15],[48].

The research work leading to this thesis dealt with proposing, implementing and evaluating improved versions of the CFDAMA protocol that can inter-work effectively with IP-based protocols. The fundamental objectives of the research are as highlighted below:

i. To design a new predictive CFDAMA protocol that can provide significant enhancements to the prediction efficiency of the existing predictive CFDAMA protocol in the literature.

ii. To implement optimizations of the CFDAMA protocol variants by exploiting the classifications of the applications in a multi-class UT application system using the application layer information provided by RTP/RTCP for capacity allocation in the CFDAMA protocol.

iii. To design an associated traffic-scheduling protocol to effectively manage and control packet access to the assigned capacity in each UT with the aid of the application layer information service provided by RTP.

iv. To evaluate the overall performance of GEO BSA system with the integration of the designed MAC protocol and traffic scheduling protocol.

Chapter 2 Modeling MAC protocols for broadband satellite networks

2.0 Introduction

In this chapter, we present an overview of BSN modeling from the viewpoint of the QoS provisioning capability of MAC protocols. In discussing the existing MAC protocols, we explore the important relationships between the applications and services that can be supported in the UT's in a BSN and the MAC protocols; and develop categories of applications services and UT types as well as performance features of MAC protocols that suit the various categories of UT types developed. We also provide a brief outlook on the performance metrics for comparing MAC protocols.

2.1 MAC protocols and QoS provisioning

In a satellite environment, MAC protocols allow the statistical multiplexing of the packet traffic from the various UT's in the return channel of a satellite to achieve the desired performance objectives at varying levels of capacity utilization. However, depending on the system environment, the MAC protocol technique can also have a dominant effect on the ability of the system to deliver the needed traffic QoS guarantee. The desired features of any MAC protocol in a satellite system are: QoS provisioning, efficient capacity utilization and service interoperability [14].

The QoS provisioning ability of a MAC protocol implies it must be able to guarantee packet-level QoS contract specified at the beginning of a communication session. At the MAC sub-layer, QoS guarantees are described by statistical guarantees on packet delay, packet delay variation (jitter) and packet loss. The MAC protocol efficiency is defined in its ability to maximize capacity resource usage while guaranteeing QoS of admitted packet flows. The service interoperability property means that the protocol must seamlessly support the same services standardized by the ATM forum or the Internet Engineering Task Force (IETF) [14]. These desired properties are often conflicting in many applications and thus pose many challenging design issues. Moreover, BSN's are designed to typically support many applications with wide ranging MAC sub-layer QoS requirements. The following is a list of current and emerging applications of BSN systems:

i. Interactive computing, e.g., Telnet applications, distributed simulations, distributed control system, which are sensitive to delays and delay variations;

ii. Bulk data transfer applications: are applications in which the total amount of data to be transferred is known a priori. They are typically not sensitive to delay or delay variations but highly sensitive to packet losses. They rely on TCP for reliable transfer;

iii. Video broadcasts applications: the broadcast property of the downlink channel of satellites enhances the delivery of these kinds of applications via BSN's. They are more sensitive to delay variations. Examples of applications in this category includes video on-demand applications, live video streaming applications;

iv. Video conferencing applications: the requirements of this class of applications depend on the desired QoS (which is related to the desired video quality) in terms of delays and delay variations and packet

The typical scenario is that the traffic generated at the UT consists of a combination of these application types. Since each application require QoS guarantees that are peculiar to their traffic characteristics, the QoS provisioning functionality of MAC protocols have to be sensitive to the variations in the traffic types. Thus, in modeling a BSN, in addition to a suitable application model, there is the necessity to define and model the service categories that can adequately characterize the properties of the several applications that can be supported in a BSN. Thus, the service model for applications in the system will be defined based on the application model. Figure 2.1 shows a dependence diagram to present the overall model of a BSN in terms of QoS provisioning in the MAC sub-layer.

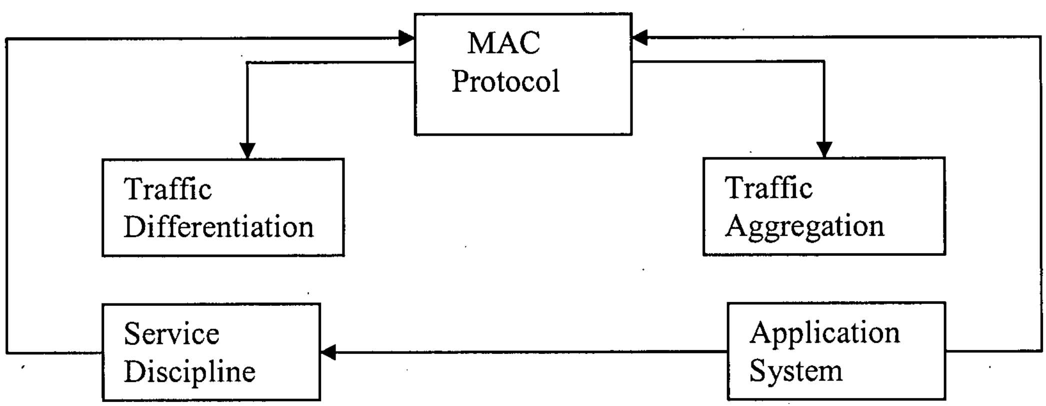

Figure 2.1: Relationships between the MAC protocol model, the application model and the service model

2.2 Application modeling

The application model typically describes and characterizes the important features of particular end-user application in a UT. The model captures statistical properties of the real application such as probability density functions (PDF) of packet sizes, packet inter-arrival times and communication session durations. Typical examples of these models include the following.

2.2.1 Voice application model

The voice application models vary widely with the characteristics of the encoding algorithm. A common model for voice applications used in simulation-based analyses is the exponential ON-OFF model [17], [18],[19]. The ON-OFF states model the talk spurts and silences in voice conversations. In the ON state, the model generates fixed-size (constant bit rate CBR) or variable-size (variable bit rate VBR) voice packets at periodic intervals defined by the sampling rate of the encoder. In the OFF state, no packets are generated. The ON duration is assumed to be exponentially distributed with a mean of 0.352 seconds while the OFF duration is taken to follow a hyper-exponential distribution consisting of a lower exponential distribution to model inter-word silences and the upper exponential distribution to model listening silences. The model also includes probabilities of the voice packet generation mechanism being in each of the distributions. Other models of voice applications include characterizations of silence estimation algorithms for reducing voice application data-rates [19].

2.2.2 Video application model

The most common video applications in today's networks are video streaming applications and interactive video conferencing applications. In general, the models for video applications for simulation-based analyses can be categorized into two classes: scene/encoder dependent models and scene/encoder independent models. Scene/encoder dependent models rely on the features of the application itself (e.g. video conferencing scene properties and motion pictures scene properties of video streams) as well as the encoder standards (H.261, H.263. H.264 or MPEG-2/4) in defining the statistical/stochastic characteristics of the models describing the video applications [20 - 22]. In [28], statistical analyses of several MBone H.261 video traces based on three encoding parameters: bit rate (B), frame rate (F) and quality (Q) provided useful results to characterize the statistical behaviors of H.261-type video applications.

A description of the strategies for modeling MPEG-2/4 video applications in terms of an empirical statistical distribution for the frame size, the frame rate and a Markov chain to characterize the burstiness in the frame size (consistent with changes in scene activity level) was given in [22-23]. Some other models of video applications are based on the distributions of the group-of-pictures (GOP) property of MPEG video rather than per-frame distributions [24-25].

Several stochastic models are presented in [26-27] all of which apply to specific kinds of video encoding standards. With the recent availability of large data sets of actual variable bit-rate (VBR) traffic measurements, many video source models that are scene/encoder independent are being developed. Several statistical analyses of these large data sets have revealed an inherent self-similarity or long-range dependence (LRD) characteristic of video sources in contrast to short range dependence (SRD) characteristic of many well-known models [34]. The fundamental behavior of LRD traffic models is primarily described by burstiness on a wide range of time scales [30-33].

The statistical analyses conducted in [32] highlighted the relationships between the burstiness in several video traffic traces and the activity levels in the traces that are intrinsically related to scene-change characteristics. The activity levels were also characterized by a statistical parameter called the *Hurst* parameter (*H*) [32]. This new modeling paradigm, which is now well known to be scene/encoder independent, allows the development of a more universal model for VBR video applications thus allowing robust comparisons of VBR video transport capabilities of various protocols and networks to be undertaken.

2.2.3 Data application model

Statistical analyses of several high-resolution traffic traces of LAN/WAN traffic have also shown the existence burtsiness over a wide range of time scales in the traffic characteristics of data application protocols [34-36]. Hence, many current models of data application protocols (or traffic sources) such as file transfer protocol (FTP), hyper-text transfer protocol (HTTP) that are being used in simulation analyses are based on the notion of self-similarity or LRD as opposed to previously used SRD models such as the exponential ON-OFF processes or Poisson traffic models. Details of the causes of the observed LRD characteristic in data applications, its statistical characteristics and its implications on network performance have been reviewed in [37].

In the satellite environment, the inherent bursty characteristics of a wide range of UT's makes the use of self-similar traffic sources in the modeling and simulation of BSN's a necessary technique. The existence of diverse user-traffic types with diverse QoS requirements also necessitates an evaluation of the impacts of self-similar data applications sources in the control and delivery of a diverse QoS contract via the MAC protocols in BSS's. The common traffic models for data applications used in simulation-based studies include the superposition of alternating ON-OFF process with heavy-tailed distributions for either the ON state or OFF state or both [15], fractional Gaussian noise (FGN) or fractional Brownian motion (FBM) models and doubly stochastic Poisson processes (DSPP) [38].

2.3 Service modeling in a BSN

In a BSN with a shared return channel, the MAC protocol implements the service discipline in the return channel. The service discipline influences how packets from various sources are transmitted in the network. The characteristics of the service disciplines are typically specified (quantitatively or statistically) in terms of some QoS-related variables such as packet delay, packet delay jitter, throughput or some predefined access priority to network resources.

The multi-service nature of the end-user applications in a typical BSS requires the implementation of diverse service disciplines to meet the requirements of different applications. Thus, the service models that can be developed in a BSN will rely on the characteristics of the models for the applications in the UT's. One of the desired performance objectives of a MAC protocol is to ensure interoperability among the different network protocol technologies (ATM-based and IP-based) in order to ensure adequate inter-working between the two categories of networking technologies. For multi-service applications in ATM networks, there are well defined service categories to implement the service disciplines for the various application types that can be supported in the network [1],[12],[13]. They are defined at each node along the route between the traffic source and the intended destination in the ATM network.

For multi-service applications in IP-based networks, the integrated services (IntServ) together with the resource reservation protocol (RSVP) specify three service categories for Internet applications: guaranteed service (GS), controlled load (CL) and the best effort service [12]. In addition to these service discipline frameworks, IP-based networks have the differentiated services (DiffServ) protocol framework that specifies the categories of services for traffic types based on per-hop behaviors (PHBs) of the aggregate flows of packets streams in the network [40]. The PHBs (expedited forwarding (EF), assured forwarding (AF) and default (DE)) determine how packets are forwarded from one node to another in a terrestrial network.

For IP-based satellite systems to successfully compete with terrestrial network systems, techniques for the service categories developed for applications in the IP-based satellite environment to inter-work seamlessly with those currently deployed in terrestrial systems need to be developed.

2.3.1 Models of application service categories for a BSN

In [2], several categories of services that can be implemented in a MAC protocol for multi-service applications support are proposed. Techniques integrating a satellite MAC protocol and current DVB-RCS open standards with DiffServ service disciplines have also been proposed in [2],[3]. For a BSN to support a broad range of application services and to enhance inter-operability of the possible service categories of a satellite MAC protocol with the existing service disciplines of IP-based protocols (IntServ and DiffServ) as well as the ATM-based service categories, we propose the following range of amalgamated service categories for a satellite MAC protocol.

2.3.1.1 Real-time variable bit rate/Guaranteed service/Expedited forwarding (rt-VBR/GS/EF)

These service categories are logically identical. The characteristics of the service implemented by all of these categories include strict bounds on packet delay, packet delay jitter and packet loss. They are effectively designed for real-time applications such as voice and video applications.

2.3.1.2 Non-real-time variable bit rate (nrt-VBR)

This service category can be applied for applications that require strict bounds on packet loss with very relaxed delay or delay jitter bounds [12].

2.3.1.3 Available bit rate/Controlled load/Assured forwarding (ABR/CL/AF)

The specifications of these service categories are logically identical. Available bit rate service is meant to give QoS guarantees that can possibly change over a life of a connection. The controlled load service category provides a similar service contract [12]. The AF PHB in DiffServ allows traffic from applications to be classified into four different levels with three levels of drop precedence [40]. The overall characteristics of this amalgamated group of service categories are light packet-loss tolerance and packet delay or delay jitter requirements that are not as strict as the ones of rt-VBR or GS.

2.3.1.4 Unspecified bit rate/Best effort/ Default (UBR/BE/DE)

These service categories are also identical in that they are meant to support applications with no critical delay, jitter or packet loss requirements. They can be applied as the default categories for the nrtVBR service category particularly for applications with elastic transport-layer flow and congestion control mechanisms where packet losses can be adequately compensated for via the packet delivery feedback mechanisms that are present in the transport protocols. Given the application model features and the service discipline model, the characteristics of a UT in a BSN can then be modeled. We now present how the models of service disciplines and the application traffic influence the choice of UT models and MAC protocols.

2.3 User terminal modeling

In a BSN, the UT's serve as the access points (to the satellite network) for end-user applications by performing a number of signaling functions by which access to the satellite channel capacity is controlled via the MAC protocol. In the MAC sub-layer, the UT's can also perform a number of actions such as error detection and correction to ensure the MAC protocol's stability [14]. It is also possible for the UT to perform a local scheduling of the possibly diverse application types supported in the end-user system as a part of achieving the overall end-to-end applications' QoS requirements implemented in the MAC protocol. Hence, in modeling BSN's, the UT model forms an essential part of the overall MAC protocol model. As shown in the dependence diagram of Figure 2.1, the characteristics of any UT model in a BSN will significantly depend on the types of applications to be supported and the service categories designed to support the applications.

Two traffic-centered paradigms can be applied in developing the models of UT's. These are:

i. Traffic differentiation;

ii. Traffic aggregation.

These paradigms are related to the need to individually model the traffic behaviors of the various traffic types in the UT's and the desire to model the effective traffic behavior of a UT.

2.3.1 Traffic differentiation

The traffic differentiation paradigm that we are introducing is similar to the underlying principles of the DiffServ framework. In this framework, the traffic streams from the various source types in the UT's can be organized into various flow types based on the source type and the characteristics of the service disciplines they require. Each flow type can thus be thought of as made up of several individual sources many of which will typically require similar service requirements from a MAC protocol. Thus, the MAC protocol functions such as RRS, DCA as well as packet transmission scheduling functions can be based on the observed characteristics of the individual traffic sources which are separately modeled. Based on the differentiation principle, the UT's in a BSN can be generally grouped into three categories:

i. Strictly real-time traffic (SRTT) UT;

ii. Strictly non-real-time traffic (SNRTT) UT;

ii. Partly real-time/non-real-time traffic (PRNRTT) UT.

A SRTT UT model will support applications with purely real-time (RT) QoS requirements while a SNRTT UT model would support applications with non-real-time (NRT) service requirements. A PRNRTT UT would have a hybrid of applications some of which will require RT QoS requirements and some NRT requirements. These classifications combine the characteristics of the applications supported in the UT's and the overall service requirements of the applications and they can be integrated to influence the DCA and RRS and TS functions in a MAC protocol.

2.3.2 Traffic aggregation

The use of traffic aggregation in modeling a UT behavior has been illustrated in [12]. The central idea in the traffic aggregation modeling paradigm is the description of the UT's in a broadband system by an effective traffic source with an offered traffic equal to the aggregate traffic from sources that are either physically or logically connected to the UT. The aggregate flow can then be described by an adequate model with traffic characteristics (packets size PDF, inter-arrival time distributions) depending on the degree of traffic stream aggregation in the UT's.

The following describes the three UT model types resulting from the traffic aggregation principle in the UT's [12]:

i. Very high source aggregation (VGSA) UT model;

ii. High source aggregation (HSA) UT model;

iii. Medium source aggregation (MSA) UT model;

2.3.2.1 Very high source aggregation (VHSA) UT model

A VHSA UT model is used to characterize the effective traffic source in large earth stations supporting a large number of statistically multiplexed traffic sources. The traffic sources can be physically or logically connected to the UT. At very short time scales, the effective traffic might be bursty but the traffic will generally be effectively smooth on a wide range of time scales due to the high statistical multiplexing. This UT model is typically appropriate to characterize the traffic at the GES and UT's supporting large scale Internet service provider (ISP) traffic requiring high bandwidth uplink connections with a delay close to propagation time [12].

2.3.2.2 High source aggregation (HSA) UT model

The HSA UT model characterizes the aggregate traffic behavior in a UT supporting statistically multiplexed traffic sources with the degree of aggregation much less than the very high source aggregation scenario. The desired bit rates are in the range of 10-100Mb/s [12]. Hence, the typical application of this UT model is in the simulation analyses of inter-connected corporate LAN's via satellite links. Depending on the degree of source aggregation, this UT model can also be used to characterize the aggregate traffic situation in the connection between an ISP and the terrestrial Internet. The basic characteristic of this UT model is a significant level of burtsiness in the offered traffic.

2.3.2.3 Medium source aggregation (MSA) UT model

The MSA UT model is employed in characterizing the behavior of the aggregate traffic in a UT with a bit rate of below 10Mb/s [12]. The MSA scenario is typically employed in modeling the traffic behavior of UT's supporting personal communications applications demands of residential, small-office-home-office (SOHO) or small corporate LAN's which are all characterized by a high degree of traffic burtsiness thus requiring dynamic MAC protocols for efficient capacity utilization.

2.4 MAC protocol models

In a BSN, the choice of a MAC protocol for a particular application and the algorithmic design of the MAC protocol for the application are both influenced by the characteristics of the UT. In the previous section, we presented a review of the characteristics of UT models by emphasizing two notions of UT modeling (differentiation and aggregation) and developed categories of UT's based on these notions. We now present the relationships between the categories of UT's defined under both traffic differentiation and aggregation principles and the MAC protocol models that have been developed in the literature. We organize the MAC protocols into the FAMA, DAMA, RA and hybrid protocols.

FAMA protocols are most effective for networks composed of small number of stations with stable and predictable traffic patterns. The pre-assigned capacities allows for the delivery of strong QoS guarantees [14]. From the characteristics of the UT model categories earlier developed, it can be asserted that the FAMA protocols can satisfy the QoS requirements of the SRTT UT category of terminals under the differentiation principle and the VHSA category of UT's under the aggregation principle. The strictly real-time characteristic of the traffic in a SRTT UT category of terminals necessitates the use of MAC protocols that will effectively guarantee minimal delays and delay jitter. FAMA protocols can deliver this requirement because they incur no delay overhead due to request reservation. Additionally, in the FA MAC protocols, the pre-assigned capacity can be guaranteed to be sufficient for the traffic emanating from the UT's. This is at the expense of capacity utilization efficiency when the UT traffic demands are low compared with the pre-assigned capacity. Similarly, the traffic characteristics of the VHSA scenario requires the use of FAMA protocols to guarantee the availability of high bit rates and low delays for the typically smooth or predictable traffic effectively generated in the UT's.

However, oftentimes, the VBR characteristics of the traffic generated in these UT's can result in significantly poor link utilization [14]. Hence, for many situations in which the traffic characteristics are well known to be bursty and with less stringent delay and delay-jitter requirements, the DAMA protocols are employed. A DAMA protocol consist of a phase for capacity requests (request reservation phase), in which UT's transmit their instantaneous capacity requirements to the capacity allocation algorithm of the protocol, and a data transmission phase in which the data packets in the terminal queues are transmitted according to the share of the channel capacity assigned each UT. The reservation phase of a DAMA protocol results in a delay overhead. Due to the limited capacity in the channel, in some instances, the share of the capacity assigned to some UT's may be less than the capacity requests. Hence, DAMA protocols cannot provide as strong QoS guarantees as can be provided by FA MAC protocols.

However, DAMA protocols can allow for a flexible control of QoS requirements through priority-based assignment techniques [14],[42]. Based on these features of DAMA protocols, it can be asserted that the requirements of the SNRTT UT can be effectively met with the use of DAMA protocols since the UT model do not have stringent delay or delay jitter requirements and have inherently bursty traffic situations. Similarly, if there is no a stringent delay or delay-jitter requirements, DAMA protocols can adequately meet the requirements of the UT traffic in the MSA UT models.

RA MAC protocols [14] are typically appropriate for networks composed of bursty traffic sources. They have the advantage of ease of implementation and since there is no central control, signaling for channel access and algorithmic processing overhead are not incurred. However, they do not provide any QoS guarantees because collision-free capacity reservations are not possible [14]. With adequate flow control mechanisms (e.g., TCP flow control mechanisms) in the higher-layer protocols and depending on the delay or delay-jitter requirements of the UT applications, RA protocols can be employed for the MSA and SNRTT UT models.

Hybrid MAC protocol models attempt to harness the desirable features of the FAMA, DAMA and RA MAC protocols. For example a hybrid FAMA-DAMA protocol typically divides the channel capacity into two segments, one of which is statically assigned to the UT's and the other assigned on a dynamic demand-assignment basis. Thus a strong upper bound on QoS parameters can be provided while efficient capacity utilization can also be ensured via dynamic allocations of some of the channel capacity. This class of hybrid protocols could effectively provide the QoS requirements of the PRNRTT UT and HSA UT models. In the PRNRTT UT case, the UT's with the real-time requirements are statically assigned some fixed capacity while the UT's with non-real-time requirements are dynamically assigned capacity. In the HSA case, all the UT's can be assigned some fixed capacity while the remaining capacity can be dynamically assigned based on the traffic situations in the UT's.

A RA-DAMA hybrid protocol also combines the desirable features of RA MAC and DAMA protocols. This hybrid protocol attempts to achieve lower end-to-end delays at the expense of light-packet loss by assigning the channel capacity in a demand-assigned manner and the remaining capacity is assessed using random access [14]. In [14], the simulation results presented on the performances of the FAMA, DAMA and the hybrid protocols (FAMA-DAMA, RA-DAMA protocols) for a medium bit rate video traffic showed that the

DAMA protocols could provide a superior performance when compared with the FAMA protocols and the hybrid protocols in terms of delay and delay variation at high network load.

Recent MAC protocol analyses for BSN's have also shown the superiority of a hybrid protocol called the combined-free demand assigned multiple access (CFDAMA) protocol [13],[15],[16],[41]. The CFDAMA protocol re-assigns the unused capacity in the channel to the UT's after the initial assignments have been made via FA or DA. Several simulation results presented in [12] and [13] show that CFDAMA offers could provide a lower delay and delay jitter at a high network load when compared with all other hybrid protocols or the FA and DA MAC protocols.

Several capacity requests schemes for CFDAMA have also been evaluated in [16], [41]. Hence, for BSN's with a typically large number of UT's (high network load) each supporting a number of bursty VBR applications, CFDAMA can be employed to provide the essential requirements of minimal delays and delay jitter at this high network load. CFDAMA can also allow for the needed flexibility in the distribution of the free capacity to satisfy the diverse QoS requirements of the applications in a UT [3]. In following chapters, we provide performance analysis of the CFDAMA protocol models and additionally provide algorithmic enhancements to the CFDAMA protocol to fully harness its potentials for QoS provisioning in a BSN. Table 2.1 below presents a relationship between traffic models and MAC protocol choices.

Table 2.1: The relations between traffic models and MAC choices

Traffic Model	MAC Class Choice
Non-bursty traffic	Fixed assignment
Bursty traffic , short messages	Random access
Bursty traffic, long messages and large number of users	Reservation protocols with contention
Bursty traffic, long messages and small number of users	Reservation protocols with fixed TDMA reservation channel

2.5 Performance comparisons of MAC protocols

In addition to the fundamental architectural objectives such as channel stability, protocol scalability, channel reconfigurability and low complexity of the control algorithm that influence the design of MAC protocols for satellite communications, the end-to-end packet delay and delay-jitter as well as the channel throughput as functions of the network traffic load represent the performance metrics for comparing MAC protocols. The packet delay metric typically combines the access delay experienced in the TDMA frame, the queuing delay experienced by packets in the UT queues and the propagation delay typical of satellite links.

Table 2.2: Performance comparison of some common classes of MAC protocols

MAC Protocol	Average Throughput	Mean delay	Stability	Scalability	Recon-figurability	Broadband applications
Fixed Assignment: B-TDMA G-TDMA	 Low High	 Low-Med Low	 Med-High High	 No No	 No No	 Yes yes
Demand Assignment	Med-High	Med-high	Med-High	No	No	Yes
Random Access (S-Aloha)	Low	Very Low	Low	Yes	Yes	No
Reservation (R-Aloha)	High	Very Low	Med	Yes	Yes	No
Hybrid Aloha-R	High	Low-Med	Med	Yes	Yes	Yes

Table 2.2 above presents a list of FAMA, DAMA, RA and hybrid protocols and their performance characteristics. For BSN systems with multi-type applications in the UT's, an additional performance metric will be flexibility in the delivery of the QoS guarantees for the different application classes. We shall explore this additional performance metric in our objective of enhancing the CFDAMA protocol algorithm for multiservice applications.

2.6 Summary

In this chapter, we have reviewed the concepts for modeling BSN's from a QoS provisioning point of view by discussing a hierarchical method for modeling MAC protocols. We have shown how the models of applications, services and UT's together with the relationships between models in the hierarchy influence the choice of MAC protocols for different applications. We have developed classifications of UT that will useful in determining the choice of MAC protocols and the MAC protocol algorithm for different applications and services.

However, since the main focus of this research is on optimizing the performance of the CFDAMA protocol for the BSA system, we briefly highlighted the potentials of the CFDAMA to provide superior performance metrics at high network loads. In the following chapters, we provide detailed algorithmic description of the CFDAMA protocol and analyze techniques and methods (relating to the protocol) that have been developed in the literature.

Chapter 3 Empirical analysis of the CFDAMA protocol

3.0 Introduction

This chapter presents the performance analyses of the prominent variations of the CFDAMA protocol. To present the algorithmic overview of the DCA function of the protocols, an empirical framework for the performance analyses in a BSA system composed of UT's with bursty VBR traffic sources is developed. The developed framework led to performance characterizations of the CFDAMA protocols based on a number of dynamic free capacity assignment scenarios that are implicitly related to the overall packet delay and jitter performances of the DCA schedulers of each of the protocol variations. Our initial analyses and the ensuing performance statistics were validated via simulations. The results provide comparisons of the average packet delay and delay jitter of the CFDAMA protocol variations as functions of the network traffic load.

3.1 Algorithmic description of the CFDAMA protocol

The CFDAMA MAC protocol is a dynamic, demand-based MAC protocol designed to enhance the channel utilization by re-assigning the free slots that are often available in the data sub-frame of a TDMA or MF-TDMA frame at the low-to-medium range of channel traffic load. For UT's with VBR sources, the dynamics of the traffic offered by each UT typically varies from one time frame to another in the TDMA/MF-TDMA channel. The activity levels of the UT traffic sources are reflected by the burstiness they exhibit in terms of size (bits or bytes) and the inter-arrival time variations of the packets.

The dynamic variations in capacity demand create several issues relating to efficiency in the utilization of the assigned capacity in addition to the optimization of the critical MAC protocol performance metrics subject to the constraint of the available system capacity. The CFDAMA protocol was proposed as a means to optimize the performance of DAMA protocols by optimizing the utilization of the free capacity in the uplink/return channel as well providing optimal MAC QoS performance requirements of which packet delay and jitter form the most critical. The method employed in allocating the free capacity to the UT's is an important feature of the CFDAMA protocols. Many algorithms for implementing the CFDAMA protocol have been proposed in the literature and the performance comparisons have shown improvements in the performance of the CFDAMA protocol variants relative to the pure DAMA protocol [3],[13],[15],[16]. To provide a background to compare these protocol variants, we have devised an analytical performance framework for evaluating the DCA algorithms of the protocol variants.

1	2	3	*Control sub-frame*	n	1	2	3	*Data sub-frame*	n

Figure 3.1: A simple TDMA frame format for transmission in the uplink channel

Figure 3.1 above shows the architecture of a TDMA transmission channel frame that will be used in developing the analytical framework. As shown in the figure, the TDMA channel is divided into time frames of a fixed duration. The duration of the TDMA frame depends on the design of the satellite system and on other factors including the characteristics of the traffic supported in the system [4]. Each UT makes a request for the available slots in the data sub-frame by transmitting a control packet (which includes the request information) in its assigned slots in the control sub-frame. It is assumed that each control slot is long enough for just one control packet. However, the duration of the slots assigned to a UT in the data sub-frame may vary depending on the calculated assignment. We show this in Figure 3.1 with a varied size of assigned slots for each UT.

Packets arriving from all the sources connected to a UT are queued at the UT in accordance with a pre-defined queuing and service scheme. Once assigned some time slots in the data sub-frame, each UT transmits the data in its buffer in the data sub-frame (in the assigned time slots) in accordance with the amount of slots assigned. For the purpose of our analyses and protocol descriptions, the following are assumed for a hypothetical BSA system that will be employed in our analyses:

i. The CFDAMA MAC protocol is implemented either as a distributed algorithm in each UT or as a centralized algorithm (with the capacity allocation executed on the satellite): this limits the propagation delay between requests and allocations to one satellite round-trip-time ($SRTT$) [12],[15];

ii. The control channel employs explicit reservations for capacity requests; i.e., each UT is assigned a slot in the control sub-frame during which it makes a request for capacity based on the instantaneous size of its terminal queue at the beginning of each frame;

iii. The duration of each TDMA frame is chosen to be equal to the maximum value of the $SRTT$ in a GEO BSS [12];

iv. Each CFDAMA protocol data unit (PDU) is equal in size (bits/bytes) and one PDU can be transmitted per slot.

The following is a list of parameters to describe our analysis:

s = variable indicating the index of a TDMA/MF-TDMA time frame;

n = number of UT's in the system;

r = data-rate of the TDMA channel;

$SRTT$ = maximum value of a GEO system satellite round-trip time (approximately 270ms);

τ = frame transmission time (duration) = $SRTT$

L = total capacity (slots) in the TDMA channel;

$C(s)$ = number of slots in the control sub-frame in frame s;

$R^i(s)$ = queue size (packets) in the i^{th} UT at the beginning of frame s;

$Ar^i(s)$ = total number of packet arrivals at the i^{th} UT in frame s;

$D^i(s)$ = capacity request (slots) of the i^{th} UT at the beginning of frame s;

$A^i_{free}(s)$ = number of free slots assigned to the i^{th} UT in frame s;

$A^i(s)$ = total assigned capacity (slots) for the i^{th} UT in frame s;

$w^i(s)$ = weight function of the i^{th} UT in frame s;

$N(s)$ = number of slots in the data sub-frame of the TDMA channel in frame s;

$F(s)$ = number of free slots in the channel during frame s;

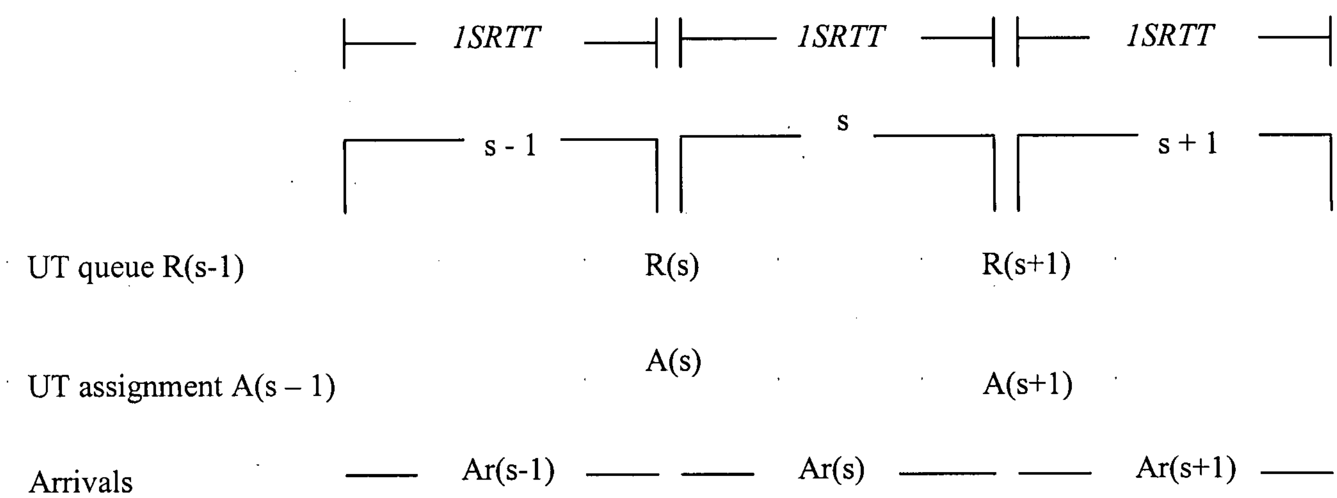

Figure 3.2: A typical variation in the UT queue

Figure 3.2 above shows the typical variations in queue status in a UT and the schematic feature of the request-assignment sequences in the CFDAMA protocol framework. In frame s, the UT is assigned capacity $A^i(s)$ which it uses to transmit the packets contributing to its queue size $R^i(s)$ and possibly, some of the packet arrivals in frame $(s-1)$. We assume that the control sub-frame is of a fixed size, that the control sub-frame is far smaller than the data sub-frame and that the request sequence for the system enables all the request packets transmitted in the control mini-slots all arrive at the UT's within one unit of the *SRTT* in the distributed system or at the satellite within one-half of *SRTT* in a centralized system. Due to the time delay between the request transmissions and their arrivals at the UT's or the satellite, the assignments scheduled for each UT in the DCA algorithm in the s^{th} frame depends on the requests from the UT's in the $(s-1)^{th}$ frame. In Figure 3.2 it is assumed that the assignment processing overhead can be neglected. This assumption, together with the earlier assumptions about the control sub-frame, allows that each UT is aware of its assignment in the s^{th} frame (for a request made in the $(s-1)^{th}$ frame) in due time for it to schedule its transmission in the s^{th} frame. The request for the s^{th} frame is calculated as:

$$D^i(s) = R^i(s) - A^i(s)$$

$$D^i(s) = 0 \text{ if } D^i(s) < 0 \tag{3.0}$$

Based on these definitions, the following relation can be written for the free capacity $F(s)$ in the TDMA channel in frame s:

$$F(s) = N(s) - \sum_{i=1}^{n} D^i(s-1) \tag{3.1}$$

Eventually, in the s^{th} frame, each UT can start its transmission in the data sub-frame according to a pre-designed transmission sequence. The size of a UT queue at the beginning of frame s is defined as:

$$R^i(s) = \left\{R^i(s-1) + Ar^i(s-1)\right\} - A^i(s-1) \tag{3.2}$$

When $A^i(s)$ is defined in frame s, we can write the residual size (which forms the demand $D^i(s)$ in frame s) of the queue in UT i as the following:

$$D^i(s) = R^i(s) - A^i(s) = \left\{R^i(s-1) - A^i(s-1)\right\} + Ar^i(s-1) - A^i(s) \tag{3.3}$$

Eqn. (3.3) shows that, depending on the variations in the packet arrival rates from the traffic sources supported by a UT and the capacity $A^i(s-1)$ assigned in the frame $(s-1)$, the values of the residual size of the queue in the UT's can vary significantly. Eqn. (3.3) is thus important in evaluating the delay characteristics of the CFDAMA protocol. If the residual queue size is large, it implies the assignment $A^i(s)$ is significantly not sufficient, and that some packets will have to be queued till the next frame $(s+1)$ before they can be transmitted. Following Eqn. (3.3), in any arbitrary frame s, an ideally optimal CFDAMA protocol can be characterized by the following argument:

The residual size of each UT queue equals zero; i.e., assigned capacity $A^i(s)$ in the frame equals the current size of the UT queue $R^i(s)$ in same frame.

Thus, $A^i(s) = R^i(s) = \{R^i(s-1) - A^i(s-1)\} + Ar^i(s-1)$

This ideal behavior results in the minimization of the average packet queuing delay and delay jitter and maximization of the average throughput on the satellite channel since it inherently describes a situation in which each UT is assigned just enough capacity to empty its queue by transmission without any residual packets in each frame. However, two factors make this ideal behavior unattainable in the system. Firstly, the capacity available in the return channel of the satellite is always limited. Secondly, in a distributed system, each UT has to estimate the arrivals $Ar^i(s-1)$ at every other UT and in the centralized DCA system, the arrivals for all the UT's have to be estimated by the payload on the satellite using specialized algorithms. These effectively result in inaccurate assignments of the channel capacity, resulting in the following realistic situations.

Situation 1: Excess capacity allocation $\left(A^i(s) >> R^i(s)\right)$

For any UT with this situation in a frame s, the packets transmitted in that frame will experience minimal queuing delays and delay jitter. However, depending on the magnitude of the difference between $A^i(s)$ and $R^i(s)$ and the potential arrival of new packets (within the allotted transmission duration) that might use the excess capacity in the UT, this situation can produce a negative effect on the throughput performance in the frame since it implies some already assigned time slots might be left unused if new arrivals in frame s cannot use those slots.

Situation 2: Insufficient capacity allocation $\left(A^i(s) \le R^i(s)\right)$

The limited capacity in the data sub-frame of the TDMA frame can result in the assigned capacity $A^i(s)$ being less than the residual queue size $R^i(s)$. This results in some packets in

UT i being delayed till the next frame period $(s+1)$ before they can be transmitted, resulting in their long queuing delays and in a wide variations in packet delays between successive frames that would eventually affect the overall packet delay-jitter performance. Both situations can occur at any UT in the system due to high traffic load and/or inaccurate assignments in the DCA algorithm. Thus, in each TDMA/MF-TDMA frame, the aggregate performance of the CFDAMA protocol will depend on the statistics of occurrences of these situations in all the UT's. However, the issue of critical importance is with the statistics of occurrences resulting from inaccurate allocations of the channel capacity.

3.2 Variations of the CFDAMA protocol

The techniques for assigning the free slots in each frame based on the underlying paradigm for estimating $A^i(s-1)$ has led to several proposed variations of the CFDAMA protocol. Each technique can be distinguished based on the definition of the UT weight function $w^i(s)$ which defines the trend in the capacity demand for each terminal. The following is a list of prominent CFDAMA protocol variations that have been researched and analyzed and in the literature:

i. The Round-Robin CDAMA (RR-CFDAMA) protocol [40];

ii. The weighted CFDAMA (W-CFDAMA) protocol [47];

iii. The predictive CFDAMA (PR-CFDAMA) protocol [15].

Let the current frame at which capacity requests are to be made be $(s-1)$. In the DCA algorithm, the following are the three cases that can occur in frame s when the assignments $A^i(s)$ are to be scheduled in the system. For all of these cases, the capacity demands $D^i(s-1)$ are given as:

$$D^i(s-1) = R^i(s-1) - A^i(s-1)$$

Case I: Total capacity request exceeds the available capacity (No free capacity)

$$F(s) = N(s) - \sum_{i=1}^{n} D^i(s-1) < 0$$

In this situation, the most logical assignment strategy is to assign each UT a proportion of the available capacity equal to the weight of its request defined as $\dfrac{D^i(s-1)}{\sum_{i=1}^{n} D^i(s-1)}$

The following expressions also hold for case I:

$$A^i_{free}(s) = 0$$

$$A^i(s) = \frac{D^i(s-1)}{\sum_{i=1}^{n} D^i(s-1)} N(s)$$

Case II: Total capacity request equal to the available capacity (No free capacity)

$$F(s) = N(s) - \sum_{i=1}^{n} D^i(s-1) = 0$$

For this case, each UT is assigned bandwidth equal to the magnitude of its request; i.e.,

$$A^i_{free}(s) = 0$$

$$A^i(s) = D^i(s-1)$$

Case III: Total capacity request is less than the available capacity (Free capacity available)

$$F(s) = N(s) - \sum_{i=1}^{n} D^i(s-1) > 0$$

For this case, the free capacity $F(s)$ is assigned to the UT's based on a weight function derived for each UT to reflect an on-going momentum in the dynamics of the traffic at each UT; i.e.,

$$A^i_{free}(s) = w^i(s)F(s).$$

Thus the assigned capacity is the total of the original demand $D^i(s-1)$ and the assigned free capacity as shown below:

$$A^i(s) = D^i(s-1) + A^i_{free}(s) = D^i(s-1) + w^i(s)F(s); \tag{3.4}$$

$$\sum_{i=1}^{n} A^i(s) = N(s)$$

Case I and case II are typical in the DCA algorithm of a CFDAMA protocol at high traffic load where many of the UT's are actively receiving packets from the application sources. Case III is prevalent in the return channel of the system in the low-medium range of traffic load and around the lower end of a high traffic load situation. Case III is important in evaluating the methods on which the derivations of $w^i(s)$ for allocating the free capacity to achieve optimum channel throughput, packet delay and jitter in the system are based. Thus to characterize the performances of the DCA algorithms in the identified CFDAMA protocol variations, we henceforth focus our analyses on case III in the following discussions.

By substituting (3.4) in (3.3), we obtain the following expressions for case III:

$$R^i(s) - A^i(s) = \{R^i(s-1) - A^i(s-1)\} + Ar^i(s-1) - \{D^i(s-1) + A^i_{free}(s)\}$$

If $D^i(s-1) = \{R^i(s-1) - A^i(s-1)\} < 0$, then the excess capacity from $A^i(s-1)$ can be utilized by some or all of the packet arrivals contributing to $Ar^i(s-1)$. Otherwise if $D^i(s-1) = \{R^i(s-1) - A^i(s-1)\} > 0$, the arrivals contributing to $Ar^i(s-1)$ will have to be delayed till the next frame s before they can be transmitted. Thus, in general:

$$R^i(s) - A^i(s) = \{Ar^i(s-1) - A^i_{free}(s)\} \quad \text{if } D^i(s-1) \geq 0 \text{ or}$$

$$R^i(s) - A^i(s) = \{Ar^i(s-1) - \{A^i(s-1) - R^i(s-1)\}\} - A^i_{free}(s) \quad \text{if } D^i(s-1) < 0 \tag{3.5}$$

Eqn. (3.5), which holds for the specific situation in which there is free capacity available in the DCA algorithm, analytically shows how the interaction of $Ar^i(s-1)$, $A^i(s-1)$ and $A^i_{free}(s)$ can influence the delay and jitter performance of the DCA algorithm in a UT queue in each frame and in an overall transmission session.

The impact of this interaction on the performance of the CFDAMA protocol can be explained in the following ways: Firstly, (3.5) implicitly explains that to minimize the residual size of the queue, i.e., $R^i(s) - A^i(s)$, and the average queuing delay and delay jitter experienced by the packets in a UT, the allocation of the free capacity in the frame is required to sufficiently meet the additional traffic demand caused by the arrivals of packets over the interval during which each UT is oblivious of $Ar^i(s-1)$ in the other UT's. This will ensure that those arriving packets are transmitted within the same frame s thereby experiencing minimal delay.

Secondly, (3.5) also describes how the average delay and jitter performances in each frame and in an overall transmission session can be controlled by varying the free capacity assignments to the UT's. Essentially, accurate information about the magnitude of $Ar^i(s-1)$ in every UT would enhance the dynamic control of the protocol performance.

Thirdly, within their transmission slots in frames s, the system's throughput efficiency is related to the ability of each UT to utilize the allocated TDMA/MF-TDMA slots. Thus, to maximize the system's overall throughput, an accurate knowledge of the trend in arrivals is necessary in order to minimize capacity wastages since the excess allocations in a frame s cannot be used after the end of frame s. Following (3.5), an ideally optimal CFDAMA protocol has the following assignment properties:

$$A^i_{free}(s) = Ar^i(s-1) \; A^i_{free}(s) = Ar^i(s-1) \quad \text{when } D^i(s-1) \geq 0$$

$A^i_{free}(s) = \{Ar^i(s-1) - \{A^i(s-1) - R^i(s-1)\}\}$ when $D^i(s-1) < 0$

In practical situations, deviations from this ideal result are typically obtained. The deviations are again caused by the fact that the free capacity $F(s)$ in frame s is limited or because in some UT's, the inaccuracies in the DCA algorithm often resulting from the estimation of $Ar^i(s-1)$ leads to inaccurate $A^i_{free}(s)$. Hence, based on (3.5), the following free assignment scenarios results are typically obtained in any UT:

Assignment Scenario I: (Sufficient free capacity assignment)

$A^i_{free}(s) > Ar^i(s-1)$ or $A^i_{free}(s) > \{Ar^i(s-1) - \{A^i(s-1) - R^i(s-1)\}\}$

In a UT, this scenario will typically result in minimal queuing delay and delay jitter and high throughput performance per frame. However, if $A^i_{free}(s)$ is significantly larger than $Ar^i(s-1)$ or $\{Ar^i(s-1) - \{A^i(s-1) - R^i(s-1)\}\}$ and the packet arrivals in frame s cannot utilize the excess capacity, significant throughput performance reduction will occur in addition to a forced poor delay or jitter performance in the other UT's with limited capacity assignments with capacity insufficiency characterized by the following expression: $A^i_{free}(s) < Ar^i(s-1)$ or $A^i_{free}(s) < \{Ar^i(s-1) - \{A^i(s-1) - R^i(s-1)\}\}$. The delay is said to be forced on those UT's because their capacity limitations could have been as a result of inaccurate assignments by the DCA algorithm.

Assignment Scenario II: (Insufficient free capacity assignment)

$A^i_{free}(s) < Ar^i(s-1)$ or $A^i_{free}(s) < \{Ar^i(s-1) - \{A^i(s-1) - R^i(s-1)\}\}$

This scenario is more prevalent in any frame s at high traffic load. It can also occur in any UT as a result of inaccurate allocations of the free capacity resulting in forced poor delay or jitter performances. Although the system's throughput in the frame will be high due to the allocated capacity being fully utilized, the delay and delay variation per packet measured in that frame could be significantly high depending on the magnitude capacity limitation in a UT.

At medium-to-high traffic load, the availability of the free capacity can be severely limited. It is important for any CFDAMA protocol variation to minimize inaccurate capacity assignments that result in forced capacity limitation in the UT's. Achieving this requires optimizing the number of assignment scenario I while also minimizing the number of assignment scenario II over any traffic session.

3.3 Performance implications of the free assignment scenarios

The following relates to the performance indices for the CFDAMA protocol;

$Th_{put}(s)$ = the throughput performance in frame s;

$P_{delay}(s)$ = the average packet queuing delay in frame s;

$P_{jitter}(s)$ = the average packet delay jitter in frame s.

The overall performance of the CFDAMA protocol variations in each frame (in terms of throughput $Th_{put}(s)$, average packet delay $P_{delay}(s)$ and jitter $P_{jitter}(s)$) is implicitly related to the statistics of occurrences of the two scenarios (described above) in the UT's.

Let $N_p(s)$ = total number of packets transmitted from all UT's throughout a frame s;

$$Pd = \lim_{T\to\infty} \frac{\sum_{m=1}^{T} Np(m) P_{delay}(m)}{\sum_{m=1}^{T} Np(m)}$$

While $P_{delay}(m)$ measures the average delay experienced by all the packets transmitted from all the UT's throughout frame m, Pd is the overall packet delay experienced throughout the duration of observation T.

$$Pj = \lim_{T\to\infty} \frac{\sum_{m=1}^{T} Np(m) P_{jitter}(m)}{\sum_{m=1}^{T} Np(m)}$$, where $P_{jitter}(s) = \frac{\sum_{m=1}^{T} \left| P^{j} - P_{delay}(m) \right|}{Np(m)}$, and P^{j} is the packet delay of the j^{th} packet transmitted in the TDMA data sub-frame m. The overall average throughput is also defined as:

$$Th_{put} = \lim_{T\to\infty} \frac{\sum_{m=1}^{T} Thput(m)}{\sum_{m=1}^{T} m}$$ where $Thput(m)$ is the throughput value in frame m. Here, m includes all the frames in the entire transmission session of duration T.

At any traffic load situation, it is desirable for a CFDAMA protocol variation to maximize the average number of occurrences of assignment scenario I in the UT's and minimize the average number of occurrences of assignment scenario II in the UT's. Of particular interest is that a protocol variant must minimize the occurrences of these scenarios due to inaccurate assignments in any specific UT, which on the overall, result in capacity wastages, poor packet delay and delay jitter performances.

Thus, to compare the CFDAMA protocol variations, we have devised statistical balance criteria that are based on the statistics of occurrence of these capacity assignment scenarios. Later in this chapter, with simulations, we will provide a strong argument for the correlations of the average packet delay and delay jitter results and the results based on our statistical performance indices.

Let $n1(s)^{-}$ = fractional number of UT's with the sufficient-free-capacity scenario in frame s;

$n2(s)^{-}$ = fractional number of UT's with the insufficient-free-capacity scenario in frame s;

Since all protocol variants execute identically when there is no free capacity in a frame, the statistical balance criteria are devised for the situations in which the MF-TDMA/TDMA frames have free capacity. Following the discussions of the characteristics of the assignment scenarios, we propose the following criteria:

Criterion I: Sufficiency parameter criterion β

The parameter β describes (on the average) the overall sufficiency of the allocations to the UT's by measuring the relative value of the fraction of UT's with sufficient free capacity allocations as described in the assignment scenario I and the fraction of UT's with insufficient capacity allocation.

$$\lim_{T\to\infty}\left\{\frac{\sum_{m=1}^{T} n1^{-}(m)}{\sum_{m=1}^{T} m} - \frac{\sum_{m=1}^{T} n2^{-}(m)}{\sum_{m=1}^{T} m}\right\} = \beta \tag{3.6}$$

where m includes only frames with free capacity, i.e., case III.

$n1(m)^{-} + n2(m)^{-} = 1$

In (3.6), the range of values of parameter β is given as $-1 \le \beta \le 1$. Values of β in the range $0 \le \beta \le 1$ implies that on the average, the number of UT's with sufficient capacity (scenario I) is greater than the number of UT's with insufficient capacity (scenario II), a condition that will effectively minimize the overall average delay. Values of β in the range $-1 \le \beta < 0$ correspond to the condition in which there are more UT's with insufficient capacity (scenario II) than those with sufficient capacity (scenario I) which on the average results in poor delay or delay jitter performances. To comparatively evaluate the accuracy of the assignments in the

CFDAMA protocols, we propose to measure the relative amount of the inaccurately assigned capacity in the insufficient capacity scenario II by defining criterion II, the insufficiency criterion and parameter as shown below:

Criterion II: Insufficiency parameter criterion δ

Let $n(s)^+$ be defined as follows:

$n(s)^+ =$ number of UT's with $\{Ar^i(s-1) - A^i_{free}(s)\} > n_p$

where n_p is a pre-determined number of packets that can be specified as a system-wide performance-related parameter. To measure the accuracy of the assignment with respect to n_p, we propose a parameter δ in criterion II as shown below:

If $\frac{\sum_{m=1}^{T} n2(m)^-}{\sum_{m=1}^{T} m}$ is significantly high (usually the case at medium-to-high traffic load);

$$\lim_{T\to\infty} \frac{\sum_{m=1}^{T} n2^-(m)}{\sum_{m=1}^{T} m} \frac{\sum_{m=1}^{T} \frac{n^+(m)}{n2^-(m)}}{\sum_{m=1}^{T} m} = \delta \tag{3.7}$$

Parameter δ has the range of values defined as: $0 \le \delta \le 1$. It measures the average fraction of UT's experiencing the condition $\{Ar^i(s-1) - A^i_{free}(s)\} > n_p$ which indicates the severity of the insufficiency in a UT experiencing the insufficient free capacity assignment as described in assignment scenario II. Generally, high values of δ indicates severe capacity insufficiency which indicates inaccurate assignments that ultimately results in severely poor delay and delay jitter performances. By combining the characteristics of these parameters as functions of the traffic load in the CFDAMA protocol variations we can determine and compare the overall accuracy in the capacity assignments.

3.3.1 The round robin CFDAMA (RR-CFDAMA) protocol

The RR-CFDAMA protocol distributes any available free capacity in the data sub-frame to all the active UT's in the network in a round-robin manner. We assume that the round-robin assignment strategy ensures the UT's have an equal share of the free capacity. Thus, if there is free capacity in frame s, i.e. $F(s) = N(s) - \sum_{i=1}^{n} D^i(s-1) > 0$; the free capacity is assigned to the UT's according to $w^i(s)$.

For n active UT's in the system, the RR-CFDAMA weight function $w^i(s)$ is defined as:

$$w^i(s) = w^i(s+j) = \frac{1}{n} \text{ for all } j = 1,2...$$

Hence, the free allocation $A^i_{free}(s)$ for any UT is defined as;

$$A^i_{free}(s) = w^i(s)F(s) = \frac{F(s)}{n}$$

$$A^i(s) = D^i(s-1) + A^i_{free}(s) = D^i(s-1) + w^i(s)F(s) = D^i(s-1) + \frac{F(s)}{n}$$

Thus, (3.5) for the RR-CFDAMA protocol is:

$$R^i(s) - A^i(s) = \left\{ Ar^i(s-1) - \frac{F(s)}{n} \right\} \quad \text{if } D^i(s-1) \geq 0 \text{ or}$$

$$R^i(s) - A^i(s) = \left\{ Ar^i(s-1) - \left\{ A^i(s-1) - R^i(s-1) \right\} \right\} - \frac{F(s)}{n} \quad \text{if } D^i(s-1) < 0$$

3.3.2 The weighted CFDAMA (W-CFDAMA) protocol

In the W-CFDAMA protocol, the weight function for each UT in a frame s is determined from the original request information $D^i(s-1)$, provided by each UT in the control sub-frame of the $(s-1)^{th}$ frame. Thus, if there is free capacity in frame s, the free capacity is assigned to the UT's based on the weight function $w^i(s)$ defined as: $w^i(s) = \dfrac{D^i(s-1)}{\sum_{i=1}^{n} D^i(s-1)}$.

$$\text{Thus } A^i(s) = D^i(s-1) + A^i_{free}(s) = D^i(s-1) + w^i(s)F(s) = D^i(s-1) + \frac{D^i(s-1)}{\sum_{i=1}^{n} D^i(s-1)} F(s)$$

From (3.5),

$$R^i(s) - A^i(s) = \left\{ Ar^i(s-1) - \frac{D^i(s-1)}{\sum_{i=1}^{n} D^i(s-1)} F(s) \right\} \quad \text{if } D^i(s-1) \geq 0 \text{ or}$$

$$R^i(s) - A^i(s) = \{Ar^i(s-1) - \{A^i(s-1) - R^i(s-1)\}\} - \frac{D^i(s-1)}{\sum_{i=1}^{n} D^i(s-1)} F(s) \text{ if } D^i(s-1) < 0 \tag{3.8}$$

3.3.3 The predictive CFDAMA (PR-CFDAMA) protocol

The foregoing discussions on the RR-CFDAMA and the W-CFDAMA protocol have highlighted their performance deficiencies in terms of utilizing the information in the traffic characteristics of the UT's for efficient capacity utilization. The RR-CFDAMA for example, makes no use of any dynamic traffic behavior in the allocation of free capacity. The W-CFDAMA protocol attempts to provide some free capacity access control and efficient capacity utilization by making use of the request information to distribute any available free capacity to the UT's.

However, because of the characteristic long propagation delay in a satellite system, significant traffic increases are possible between the instant request information are sent and the instant that the disseminated allocation information arrive at the UT's. Additionally, for highly bursty traffic sources with high variability, the W-CFDAMA will definitely result in inaccurate assignments of the free capacity. The PR-CFDAMA protocol was proposed in [15] as an enhancement to the W-CFDAMA. Although the PR-CFDAMA is also based on the derivation of a weight function for each UT for free capacity distribution, the basic feature of the enhancement provided by the PR-CFDAMA protocol is the ability to estimate the traffic changes in each UT by estimating $Ar^i(s-1)$ for the UT's. The PR-CFDAMA algorithm allocates any free capacity based on a weight function derived from the estimated value of $R^i(s)$ for each UT.

Let: $Ar^i_{est}(s-1)$ = the estimated total size of packet arrivals in frame (s-1);

$R^i_{est}(s)$ = the estimated size of the queue in a UT at the beginning of frame s.

Thus, $R^i_{est}(s) = \{D^i(s) - A^i(s-1)\} + Ar^i_{est}(s-1)$;

Hence, the allocation of any free capacity *F(s)* by the PR-CFDAMA algorithm is given by the weight function equation below:

$$w^i(s) = \frac{R^i_{est}(s)}{\sum_{i=1}^{n} R^i_{est}(s)} F(s) \tag{3.9}$$

3.3.3.1 PR-CFDAMA traffic prediction

The PR-CFDAMA protocol algorithm demonstrates the importance of using traffic characteristics in resource allocation. It is a novel attempt to utilize burstiness of user traffic in the DCA algorithm of the CFDAMA protocol. The positive varying trends of traffic at the UT's were estimated for self-similar traffic sources known to be bursty over a wide range of time scales. The value of $R^i(s)$ for each terminal queue was treated as a time series variable [15],[48]. The formulation facilitated the use of various techniques for time series prediction [15],[43]. The method prescribed for the time series prediction was the local linear approximation (LLA). The radial basis function network (RBF) and the support vector machines (SVM) methods [15] are some other prediction methods that have been investigated.

The LLA time series prediction method is often preferred in traffic estimation or prediction because of its simplicity [15]. In a general nearest-neighbor LLA method presented in [43], a number of vectors are systematically designed from the time series samples and the vectors with the shortest Euclidean distance to that containing the most recent sample of the time series is used in estimating a future time series value. In [15] and [48], the variant of the LLA method employed for Internet traffic prediction estimates the positive varying (increment) trend of the traffic in each UT from the first and second order variations of the traffic. The positive varying trend in each UT, measured from the first-order and second order variations in the queue size, provides an approximation to the trend in packet arrivals over a frame period. The following describes the trend prediction method:

$\Delta R^i(s) = R^i(s) - R^i(s-1);$

$\Delta^2 R^i(s) = \Delta R^i(s) - \Delta R^i(s-1);$

$Ar^i_{est}(s-1) = c(s)\Delta R^i(s) + (1-c(s))\Delta^2 R^i(s)$

$Ar^i_{est}(s-1) = 0 \text{ if } Ar^i_{est}(s-1) < 0$

$\Delta R^i(s)$ measures the first-order changes in queue size and $\Delta^2 R^i(s)$ measures the second-order change UT queue size.

$c(s)$ is a weight factor that is common to all the UT's and defined as:

$$c(s) = \frac{F(s)}{n}$$

The UT weights and the free capacity assignments are then obtained as in [15]:

$$R_{est}^{i}(s) = \left\{R^{i}(s-1) - A^{i}(s-1)\right\} + Ar_{est}^{i}(s-1);$$

$$w^{i}(s) = \frac{R_{est}^{i}(s)}{\sum_{i=1}^{n} R_{est}^{i}(s)} F(s)$$

Thus $A_{free}^{i}(s) = w^{i}(s)F(s)$

$$A^{i}(s) = D^{i}(s-1) + A_{free}^{i}(s) = D^{i}(s-1) + \frac{R_{est}^{i}(s)}{\sum_{i=1}^{n} R_{est}^{i}(s)} F(s) \qquad (3.10)$$

3.4 Simulations and performance discussions

To validate the foregoing empirical analysis of the CFDAMA protocol variations, we have simulated the distributed form of the protocols for a GEO BSA system.

3.4.1 Simulation platform description and methodology

We employed the OPNET simulation platform in the simulation of a BSA system. The OPNET platform provides an event-driven method for simulating the processes that occur in the system such as packet generations and arrivals, packet queuing, packet transmission and statistics collection that allows for measurement of performance metrics such as time delay and jitter. In general, the fundamental processes occurring in a UT or traffic source are modeled using as finite state machines (FSM) consisting of states which represents the particular actions to be performed in the processes. Transitions from one state to another in a process model are triggered by what are called interrupts. Figure 3.3 shows the schematic diagram of the BSA network model that was simulated. As shown in Figure 3.3, the simulated BSA system model consists of one GES and n UT's. Each UT supports an aggregate traffic source that is modeled as a self-similar process to reproduce the burstiness characteristic of most LAN/WAN and Internet traffic. We assume that the BSA system has an adequate forward error correction scheme to mitigate bit errors over the satellite channel which might result in protocol instability and transmission collisions if the control packets are lost due to bit errors and/or the calculated assignments are erroneous due to lost packets.

Figure 3.4 shows the event-based model configuration of the UT traffic sources as modeled in the OPNET platform. For consistency of evaluation, the two types of self-similar traffic sources employed in [15] were used. These are the self-similar ON-OFF UT traffic model and the superimposed fractal renewal process (SFRP) traffic model. However, to ensure statistical independence of each source, some of the parameters of the traffic sources for both the SFRP and the ON-OFF traffic sources in each UT were randomized. For the ON-OFF traffic model, the ON-OFF states were specified as heavy-tailed Pareto distributions [15]. The Pareto distribution parameters are the location parameter k which represents the smallest possible value of the random variable generated and the shape parameter α, which determines the proportion of the probability mass that may be present in the tail of the distribution.

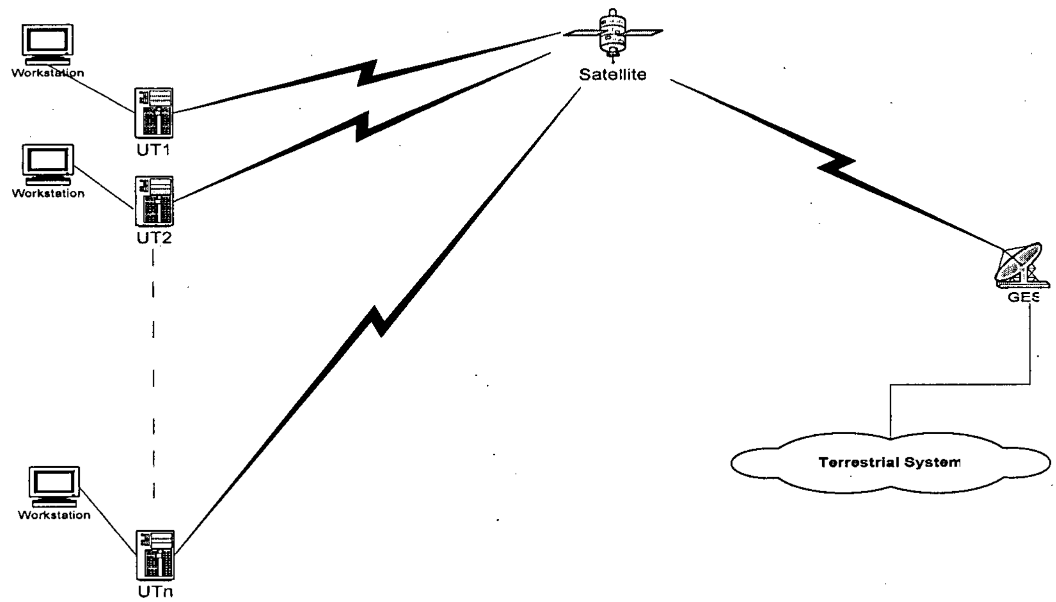

Figure 3.3: The schematic diagram of the BSA system model with one GES and n UT's.

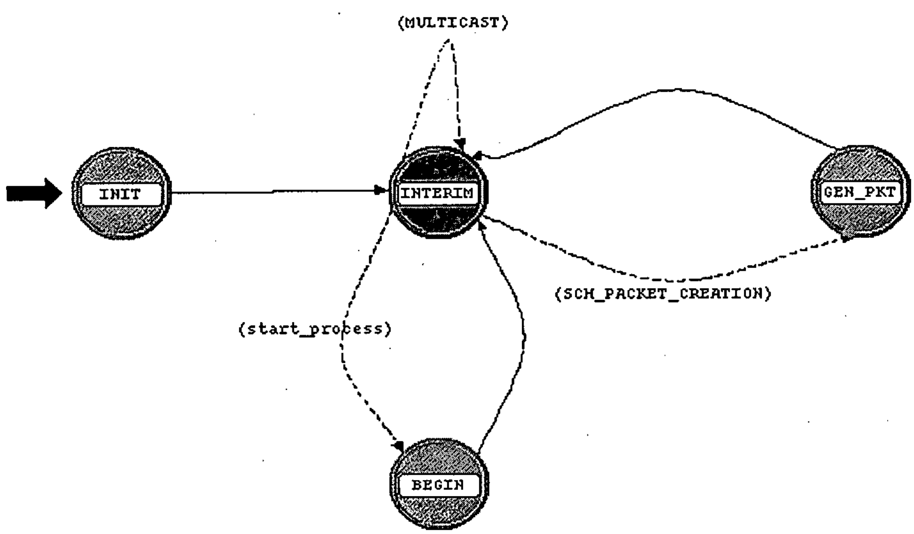

Figure 3.4: Event-based model configuration of the UT traffic sources

The location parameter k for the sources was specified to be uniformly distributed between the range (0.6-1.0) while the shape parameter α was specified as being uniformly distributed within the range (1.0-2.0) [43]. A similar randomization procedure was employed for the mean packet size distribution and the mean ON-OFF durations. This ensures that the traffic model parameters of the UT's are as different as possible to ensure as much variability as possible. For the SFRP traffic model [15], the mean length of the web-files was specified to be uniformly distributed between the range (10KB – 20KB) while the parameters describing the inter-arrival times of web traffic (which is Pareto-distributed) was specified as the same as for the ON-OFF traffic model. Tables 3.1 and 3.2 present a summary of the traffic sources' parameters.

Table 3.1: Statistical parameters for the Pareto-distributed ON-OFF traffic model [15]

Source	Parameter	Value
Uniformly distributed	Location parameter-k	(0.6 – 1.0)
Uniformly distributed	Shape parameter -α	(1.0 – 2.0)
All	*SRTT*	0.27
Uniformly distributed	Packet size (mean)	(1000- 2000)bytes
Uniformly distributed	Mean ON/OFF	(0.01-0.03)

Table 3.2: Statistical parameters for the Pareto-distributed SFRP traffic model

Source	Parameter	Value
Uniformly distributed	Location parameter-k	(0.6 - 1.0)
Uniformly distributed	Shape parameter -α	(1.0 - 2.0)
All	*SRTT*	0.27
Uniformly distributed	Mean web-file length	(10KB-20KB)

Figure 3.5 shows the internal configuration of the UT in the simulation platform. In each UT, packets arrivals are queued in the process state **collate**. Requests are made in the state **REQ** when all the UT's are notified by the MULTICAST interrupt. Based on the appropriate transmission time of the UT, the UT models initiate packet transmissions with the interrupt SELF_INTRPT_SEND in the state **start_send**.

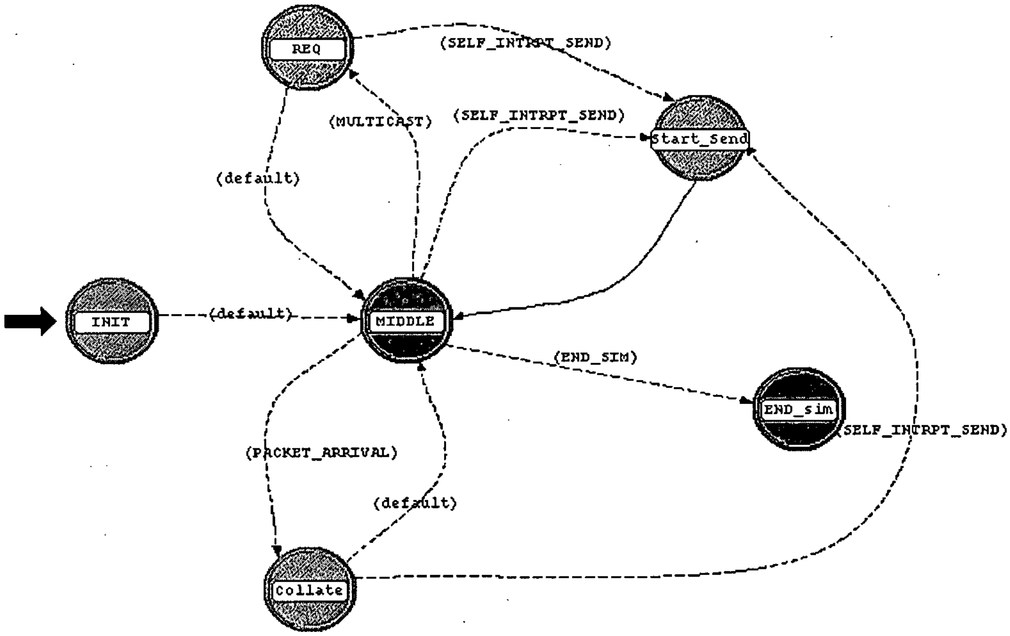

Figure 3.5: Event-based configuration of the UT in the OPNET simulation environment

3.4.2 Simulation description

UT ID	REQ_SIZE

Figure 3.6: A hypothetical structure for the control packet

UT ID	TIMESTAMP	DATA PACKET

Figure 3.7: The hypothetical structure for the PDU employed in the simulation

Figures 3.6 and 3.7 show the structure of the request packets transmitted in the control sub-frame and the structure of the PDU for the data packets respectively. We emphasize the prevalence of VBR sources with variable-length packets. Each PDU will contain a single packet. Consequently, the PDU of the protocol variants will have varying sizes. We assume that the UT's have known slots demarcations so that slots with capacity greater than a PDU can be fully utilized if it is large enough to be used by subsequent PDUs. Moreover, if the capacity of a slot is less than a PDU, then the PDU will not be transmitted in the slot. To simplify the dynamic capacity allocation, we assume that the return channel is based on the TDMA transmission mode and that the UT's transmit at the same rate. This allows for the available capacity to be distributed in terms of the number of bytes transmittable over the channel. Hence, the following relations describe the capacity:

$$L = r\tau \tag{3.11}$$

Table 3.3 presents the parameters that were employed to control the simulation.

Table 3.3: Summary of parameters to control the simulation

Simulation Parameter	Value/Range
Simulation length	10,000$SRTT$
n (Load)	$(5-100)$ UT
Seed Value	200 – 4000
r (Uplink data-rate)	4Mb/s

3.4.3 Performance results comparisons

The performance results of the protocols variants are typically compared in terms of statistical parameters such as average delay (for the system, we measure the delay between the UT and the gateway) and average delay jitter as functions of the traffic load on the return channel. The UT-to-gateway delay includes the GEO system propagation delay, the queuing delay in the UT queue, transmission delay over the return channel and the processing delay characteristic of each protocol variant. Although the superiority of the PR-CFDAMA over the RR-CFDAMA and W-CFDAMA has been confirmed in the literature, we provide additional explanations for the observed performance variations using the statistical criteria that we have formulated in our empirical analysis.

In the simulations, the traffic load was gradually increased by increasing the number of UT's in the system. The UT-to-GES average packet delay is defined as the average delay experienced by all the packets from all the UT's to the GES. The contributions to the packet delay measured in the simulations include the packet queuing delay, the protocol processing delay as well as the one-SRTT transmission delay between the UT and the GES. The delay jitter was measured relative to the average delay experienced by all the packets in the simulation.

The average traffic load T_load is defined as:

$$T\_load = \frac{\sum_{i=1}^{n} C_{av}^{i}}{r} \tag{3.12}$$

C_{av}^{i} is the offered traffic average bit-rate measured at each UT and r is the TDMA channel data-rate.

3.4.3.1 UT-to-GES average packet delay (ON-OFF Model)

Figure 3.8 below shows the average end-to-end delay between a UT and the GES for the modeled BSA system. Table 3.4 displays the upper and lower bounds of the 95% confidence interval values of the delay characteristic. At the low-to-medium range of traffic loads, the RR-CFDAMA protocol provides the best delay performance. Under these conditions, the number of UT's is small hence there are significantly large amount of free capacity available in the return channel making the requests from the UT's at the beginning of any frame s (defined as $D^{i}(s) = \{R^{i}(s) - A^{i}(s)\}$) mostly insignificant. Consequently, trend predictions in the PR-CFDAMA protocol are inaccurate since the demand information from the UT's lacks the numerical quality for effective traffic estimation. Moreover, since the free capacity is almost always equally assigned the UT's in the RR-CFDAMA protocols, the UT's are equally guaranteed access to any available free capacity in the RR-CFDAMA protocol hence its better delay performance at low-to-medium traffic load.

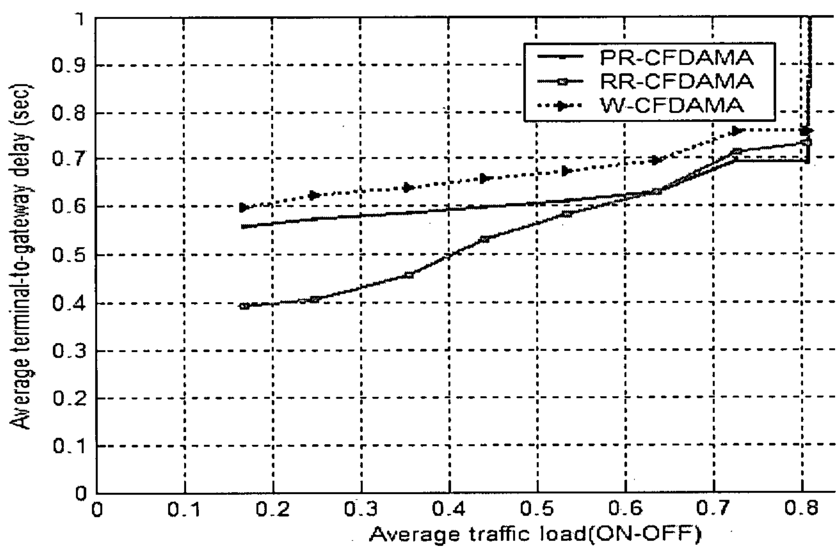

Figure 3.8: Average UT-GES packet delay as a function of average traffic load (ON-OFF)

The significance of the trend prediction mechanism in the PR-CFDAMA becomes apparent from medium-to-high traffic loads, at which $D^i(s) = (R^i(s) - A^i(s)) > 0$ thus providing the numerical quality for a more accurate prediction of the traffic trends in the UT's and a better delay performance.

Table 3.4: 95% confidence interval delay performance values for the CFDAMA protocols

	PR-CFDAMA		RR-CFDAMA		W-CFDAMA	
Load	Lower bound (sec)	Upper bound (sec)	Lower bound (sec)	Upper bound (sec)	Lower bound (sec)	Upper bound (sec)
0.1685	0.5430	0.5716	0.3891	0.3982	0.5769	0.6192
0.2490	0.5579	0.5870	0.3959	0.4121	0.6068	0.6425
0.3560	0.5699	0.5998	0.4153	0.4919	0.6213	0.6594
0.4415	0.5838	0.6119	0.4984	0.5622	0.6390	0.6736
0.5337	0.5905	0.6335	0.5015	0.6633	0.6478	0.7010
0.6354	0.6084	0.6510	0.5747	0.6840	0.6717	0.7194
0.7275	0.6102	0.9297	0.5832	0.9771	0.6860	0.9791
0.8055	0.6747	0.8097	0.5933	0.9785	0.6974	0.9870
0.9288	0.0000	21.8159	0.0000	21.6606	0.0000	21.8058

3.4.3.2 UT-to-GES average packet delay jitter (ON-OFF Model)

A similar performance trend can be observed for the packet delay-jitter performance shown in Figure 3.9. The effects of prediction mechanism in the PR-CFDAMA protocol in terms of delay jitter improvement becomes significant from medium-to-high traffic loads suggesting that the prediction accuracy of the PR-CFDAMA is significantly limited at low-to-medium traffic loads. Table 3.5 displays the 95% confidence interval values of the delay jitter performance characteristic.

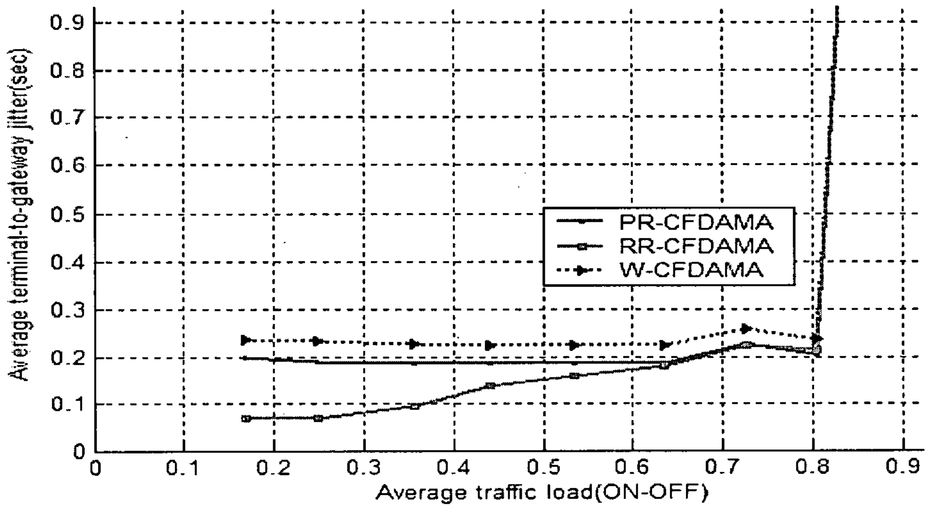

Figure 3.9: Average UT-GES packet delay-jitter as a function of average traffic load (ON-OFF)

Table 3.5: 95% confidence interval delay jitter performance values for the CFDAMA protocols

	PR-CFDAMA		RR-CFDAMA		W-CFDAMA	
Load	Lower bound (sec)	Upper bound (sec)	Lower bound (sec)	Upper bound (sec)	Lower bound (sec)	Upper bound (sec)
0.1685	0.1798	0.1903	0.0652	0.0663	0.2294	0.2395
0.2490	0.1811	0.1901	0.0664	0.0679	0.2278	0.2350
0.3560	0.1815	0.1881	0.0783	0.1051	0.2232	0.2298
0.4415	0.1819	0.1905	0.1233	0.1503	0.2215	0.2268
0.5337	0.1864	0.1938	0.1277	0.1894	0.2203	0.2271
0.6354	0.1909	0.2044	0.1701	0.1913	0.2204	0.2283
0.7275	0.0716	0.3845	0.0955	0.3542	0.2268	0.3890
0.8055	0.1361	0.2757	0.1612	0.2684	0.2873	0.3892
0.9288	0.0000	11.2872	0.0000	11.2425	0.0000	11.2611

3.4.3.3 Statistical average performance with respect to the sufficiency parameter β and insufficiency parameter δ (ON-OFF model)

The statistical average values of the sufficiency parameter β and insufficiency parameter δ are as shown in Figures 3.10 and 3.11.

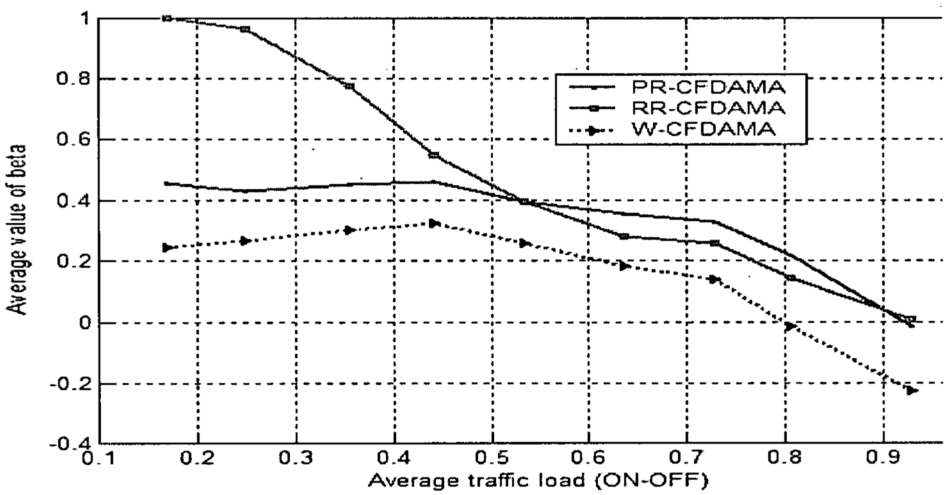

Figure 3.10: Sufficiency parameter β as a function of average traffic load (ON-OFF)

Figure 3.10 supports the observations in the delay and delay jitter performances shown in Figures 3.8 and 3.9. Typically, at the low-to-medium range of traffic load, β is significantly high, i.e., there are more UT's with sufficient capacity allocations than those with insufficient allocations, and β decreases (eventually becoming negative) as the traffic load on the return channel rises from medium to high loads. At low to medium traffic loads, RR-CFDAMA is able to guarantee more UT's with sufficient capacity allocations, while the PR-CFDAMA protocol's predication mechanism is able to guarantee more UT's with sufficient capacity allocations at medium-to-high traffic loads. This behavior can be related to the numerical quality of the demand information earlier explained for the delay and delay jitter characteristics.

Due to the fact that we considered PDU's of a variable length and because capacity requests and allocations were measured in terms of bytes of data, the insufficiency parameter δ was evaluated by setting n_p as 10 times the average size of the PDU in a frame. The value of n_p can be system-wide specified or its value can be varied for different UT's, depending on a specific UT requirement. A system-wide value of 10 for n_p specifies δ as the average fraction of UT's under an insufficient capacity allocation condition with the insufficiency (evaluated as $(\{Ar^i(s-1) - A^i_{free}(s)\} > n_p)$) greater than 10 average-sized packets. As shown in Figure 3.11, at the low-to-medium range of traffic loads, the RR-CFDAMA guarantees the minimum value of δ. As the traffic load increases beyond the medium range, the PR-

CFDAMA is able to guarantee the minimum value of δ as its prediction mechanism become more accurate in the load range.

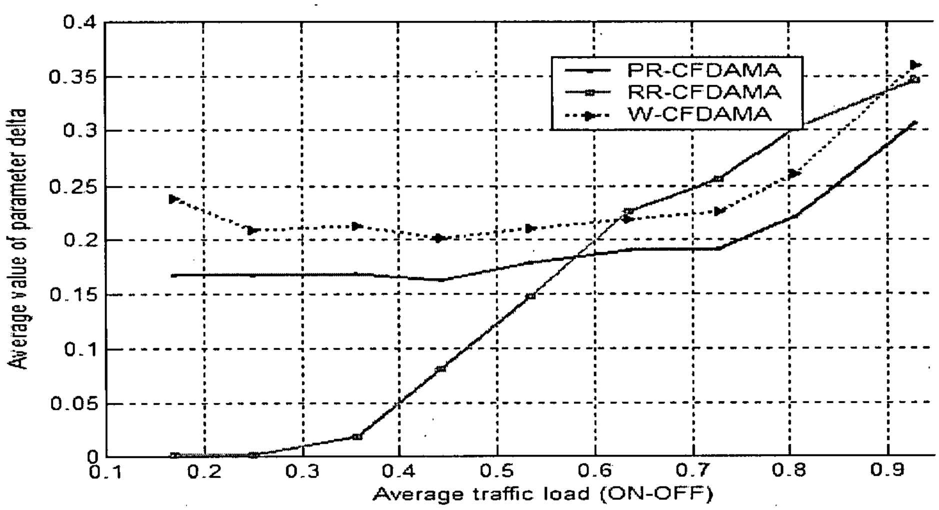

Figure 3.11: Insufficiency parameter δ as a function of average traffic load (ON-OFF)

3.4.3.4 UT-to-GES average packet delay (SFRP Model)

Figure 3.12 shows the average UT-to-GES packet delay performance for the SFRP traffic model. Contrary to the delay performance of the PR-CFDAMA protocol for the ON-OFF traffic model in Figure 3.9, Figure 3.12 shows that the PR-CFDAMA protocol has significantly better delay performance over a wide range of traffic load. The disparity can be explained by considering the higher variability (or burstiness) characteristic of the SFRP traffic model.

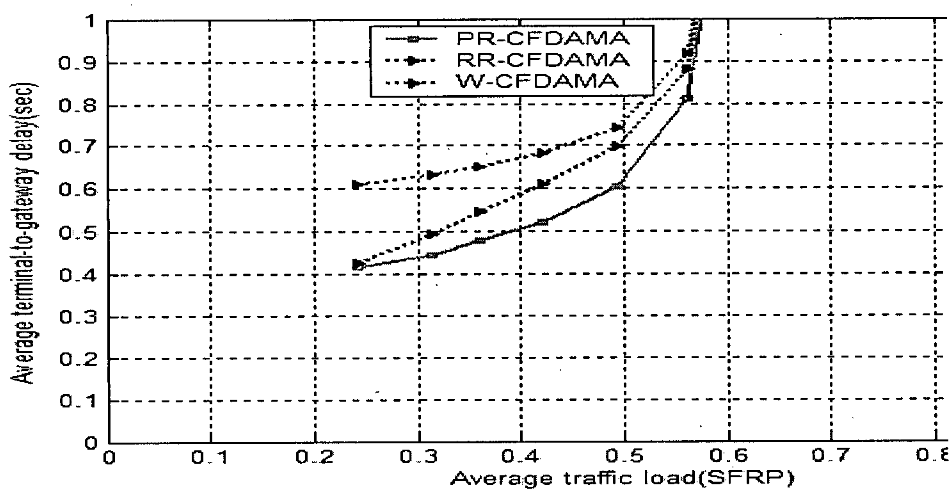

Figure 3.12: Average UT-GES packet delay as a function of average traffic load (SFRP)

At low traffic loads, the high variability allows for a significant numerical quality of the demand information even though the number of UT's is small and there is significantly large amount of free capacity. This allows for a better trend prediction in the PR-CFDAMA protocol and accounts for the better delay performance of the PR-CFDAMA protocol over the RR-CFDAMA protocol. At high traffic loads, the demand information becomes more qualitative and can accurately describe the trends in the UT traffic behavior. The PR-CFDAMA protocol still maintains its superiority in under this situation despite the fact that the capacity is more limited due to the larger number of UT's in the system.

3.4.3.5 UT-to-GES average packet delay jitter (SFRP Model)

Figure 3.13 however shows that the delay jitter performance of the PR-CFDAMA is not significantly superior to that of the RR-CFDAMA protocol due to the high variability of the SFRP traffic model. The main cause of high overall delay jitter is the delay experienced by packets between two consecutive frames. Since the RR-CFDAMA protocol allocates any free capacity in the frames to the UT's on an almost equal basis, the inter-frame component of the packet delay jitter in the RR-CFDAMA protocol can be minimal despite the high variability of the SFRP traffic.

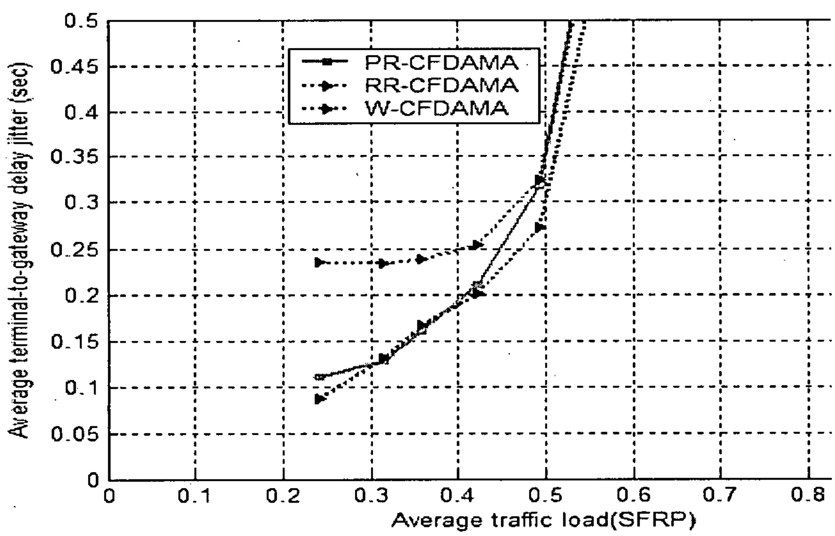

Figure 3.13: Average UT-GES packet delay jitter as a function of average traffic load (SFRP)

3.4.3.6 Statistical average performance with respect to the sufficiency parameter β and insufficiency parameter δ (SFRP model)

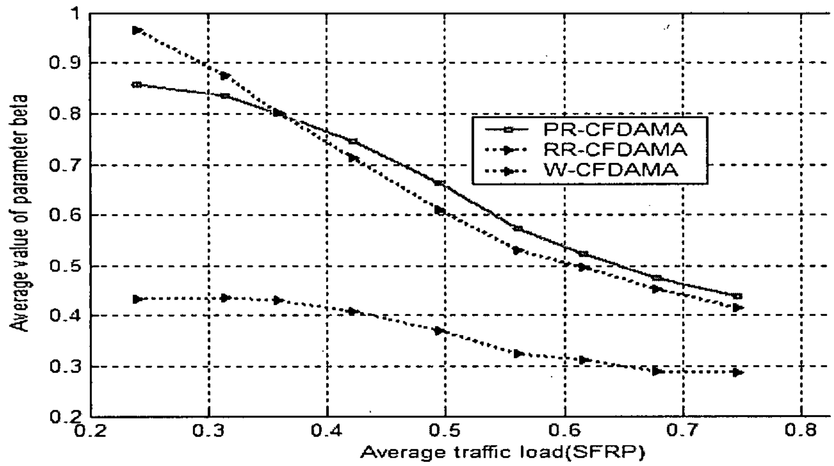

Figure 3.14: Sufficiency parameter β as a function of Average traffic load (SFRP)

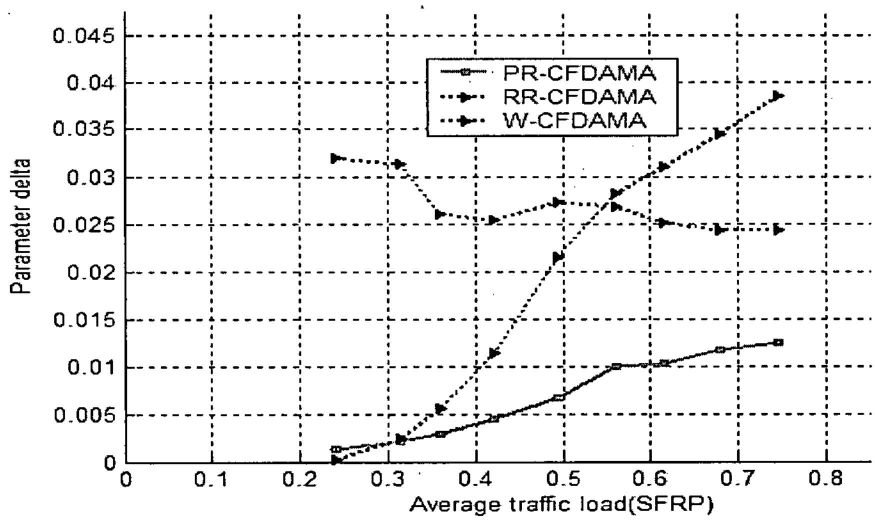

Figure 3.15: Insufficiency parameter δ as a function of the average traffic load (SFRP)

Figures 3.14 and 3.15 show that similar to the results obtained for the ON-OFF traffic model, the PR-CFDAMA protocol can guarantee dynamic assignment scenarios in the UT's that can satisfy the sufficiency criterion measured by parameter β and the insufficiency criterion measured by parameter δ better than both the W-CFDAMA and the RR-CFDAMA protocols for the SFRP traffic model.

3.5 Summary

An empirical performance analysis of the CFDAMA protocol for bursty traffic in a BSA system has been developed. The most prominent variations of the CFDAMA protocol in the literature have also been introduced and the developed analytical framework has been employed to explain their performances. This chapter has also provided supplementary statistical performance indicators that can be used to measure the system-wide performance of each of the CFDAMA protocol variations. The intrinsic relationship between these system-wide indicators and the resulting packet delay and delay jitter has been identified from the empirical analyses.

The simulation results presented for a GEO-based BSA system provided additional justifications for the superiority of the PR-CFDAMA protocol variant over the RR-CFDAMA and W-CFDAMA protocol variants under various traffic load situations. However, the simulation results also show that the PR-CFDAMA protocol performance improvement over the RR-CFDAMA protocol is mainly significant in the high system load situations. At low-to-medium traffic load scenarios, the RR-CFDAMA protocol provides better average delay and delay jitter performances over the PR-CFDAMA protocol for the ON-OFF self-similar traffic model while at high traffic loads, the PR-CFDAMA protocol's delay and delay jitter performances are superior. Moreover, for the more bursty self-similar SFRP traffic model, although the PR-CFDAMA protocol provides a better average delay performance over a wide range of traffic load situations, its delay jitter performance is not significantly better than that of the RR-CFDAMA protocol over a wide range of traffic load situations.

In the following chapter, we propose an enhancement to the PR-CFDAMA protocol that can guarantee significant improvements over the RR-CFDAMA and the PR-CFDAMA protocols.

Chapter 4 A comparative evaluation of an enhanced predictive CFDAMA DCA

4.0 Introduction

This chapter introduces an enhancement to the prediction technique for the DCA algorithm of the PR-CFDAMA protocol. The goal of the enhanced predictive algorithm is to improve the effectiveness of the estimation of the trends in packet arrivals in a UT over the RTT request-assignment delay that has been discussed in Chapter 3. As shown in the foregoing chapter, predicting the trends in the packet arrivals in the UT's during this interval is central to the overall performance of a predictive CFDAMA protocol.

In the proposed enhanced predictive algorithm, the time series of cumulative packet arrivals in two consecutive frames are used to estimate current trends in the packet generation momentum in each UT. The most recent samples of this time series are used to emulate the traffic momentum and thus estimate the trends in the actual packet generation mechanisms in the UT's. This technique contrasts to that used in the predictive algorithm for the PR-CFDAMA in which the joint dynamics of the free assignments and the packet arrivals were used in the trend prediction algorithm. The empirical logic underlying the traffic momentum technique and the comparative analyses of the predictive mechanisms in both the PR-CFDAMA protocol presented in Chapter 3 and the enhanced version (Enhanced PR-CFDAMA or EPR-CFDAMA) are discussed in this chapter.

4.1 Traffic aggregation

The busty characteristics of active (on-going) traffic in the Internet and other networks can generally be characterized both in terms of packet sizes (which indicates the number of bytes in each packet data unit) and the inter-arrival times of the packets. Thus, to vary the assignments of the free capacity in each frame, the DCA algorithm has to take into consideration these two features of UT traffic. Fortunately, in a GEO BSA system the *SRTT* can be in the range of 270ms, large variations in the packet arrivals in a UT can be tuned-down by sampling the aggregate arrivals over the *SRTT* request-assignment interval. This aggregate sampling, which conforms to our frame cycle formulation, can then form a useful basis for estimating the changes in the behavior of the traffic generation mechanism in each UT.

The residual queue size defined in (3.5) represents the number or total size of packets that will be resident in the UT queue after the assigned capacity $A^i(s)$ has been totally utilized or the excess amount of capacity that can be utilized by newly arrived packets in the current frame s. Eqn. (3.5), coupled with the request-assignment depiction given in Figure 3.2 provide the following observations:

i. To produce a performance enhancement of the predictive method presented in [15], the free assignment $A^i_{free}(s)$ must optimally meet the arrivals $Ar^i(s-1)$ in any free slots assignment in each frame. The optimality condition centers on $A^i_{free}(s)$ accurately satisfying the traffic increase caused by the new arrivals without jeopardizing the throughput performance in the system that could result if $A^i_{free}(s)$ unnecessarily exceeds $Ar^i(s-1)$. Meeting this condition necessitates accurate estimation of $Ar^i(s-1)$;

ii. In order to reinforce improvement in the delay and delay jitter performance of the predictive mechanism, the free assignment in frame s, $A^i_{free}(s)$, should not only meet the traffic increment due to $Ar^i(s-1)$ but also the potential packet arrivals in frame s. Since $A^i_{free}(s)$ will be utilized in frame s, if the arrival characteristics in frame s ($Ar^i(s)$) can be estimated in the predictive algorithm in frame s, $A^i_{free}(s)$ can be optimally assigned to satisfy all or part of the capacity requirements of potential packet arrivals contributing to $Ar^i(s)$.

For the PR-CFDAMA protocol presented in [15] and in Chapter 3, the ideally optimal inference from the first observation is to schedule $A^i_{free}(s)$ such that $R^i(s)-A^i(s)$. This effectively optimizes the delay and delay jitter performances in frame s. The second observation places a higher condition on the prediction method, implying scheduling $A^i_{free}(s)$ such that $\{R^i(s)+Ar^i(s)-A^i(s)\}=0$

To represent this analytically, (3.5) can be augmented to include $Ar^i(s)$ as shown below:

$$R^i(s)+Ar^i(s)-A^i(s)=\{Ar^i(s-1)+Ar^i(s)-A^i_{free}(s)\} \text{ if } D^i(s-1)\geq 0 \text{ or}$$

$$R^i(s)+Ar^i(s)-A^i(s)=\{Ar^i(s-1)+Ar^i(s)-\{A^i(s-1)-R^i(s-1)\}\}-A^i_{free}(s) \text{ if } D^i(s-1)<0 \tag{4.0}$$

4.2 The PR-CFDAMA protocol re-analyzed

The PR-CFDAMA protocol is based on the following equations:

$\Delta R^i(s) = R^i(s) - R^i(s-1);$

$\Delta^2 R^i(s) = \Delta R^i(s) - \Delta R^i(s-1);$

From (3.5) assuming $D^i(s-1) \geq 0$,

$$\Delta R^i(s) = \{Ar^i(s-1) - Ar^i(s-2)\} - \{A^i_{free}(s) - A^i_{free}(s-1)\} + \{A^i(s) - A^i(s-1)\}$$

$$\Delta^2 R^i(s) = \{Ar^i(s-1) - 2Ar^i(s-2) + Ar^i(s-3)\} - \{A^i_{free}(s) - 2A^i_{free}(s-1) + A^i_{free}(s-2)\}$$

$$- \{A^i(s) - 2A^i(s-1) + A^i(s-2)\}$$

$$= \{\Delta^2 Ar^i(s) - \Delta^2 A^i_{free}(s) - \Delta^2 A^i(s)\} \tag{4.1}$$

The prediction algorithm in the PR-CFDAMA protocol makes use of (4.1) to estimate the arrivals in a UT during frame $(s-1)$ according to the following:

$Ar^i_{est}(s-1) = c(s)\Delta R^i(s) + (1-c(s))\Delta^2 R^i(s)$

Eqn. (4.1) shows that PR-CFDAMA makes use of the joint dynamics of the cumulative packet arrivals of the traffic in the UT's and the dynamics of the assignments to the UT's. The main flaw of using the joint dynamics of assignments and traffic arrivals in the prediction is that the arrivals over the any frame period are independent of the assignments over any time period. Stated in another way, the traffic generation mechanisms in the UT's are independent of the capacity assigned to the UT's. Thus, while the samples of the $\Delta R^i(s)$ and $\Delta^2 R^i(s)$ contain information about the trends in actual traffic arrival variations, the traffic arrival information variables in them can be thought of as being corrupted by the assignment variables. This puts the effectiveness of the prediction algorithm of the PR-CDAMA as presented in [15] into question. In addition to this, the observation that $A^i_{free}(s)$ can affect not only packet arrivals in frame $(s-1)$ but also packet arrivals and transmission in frame s makes it necessary to use a prediction paradigm that will be based on two-frame trends rather than just one-frame arrival trend as presented in [15].

4.3 The enhanced predictive CFDAMA protocol

The enhanced predictive (EPR-CFDAMA) protocol is based on the foregoing arguments about the PR-CFDAMA protocol. To describe the analytical details of the EPR-CFDAMA protocol, (4.0) is revisited. The optimal assignment condition for (4.0) is derivable as:

$$R^i(s) + Ar^i(s) - A^i(s) = \{Ar^i(s-1) + Ar^i(s) - A^i_{free}(s)\} = 0 \quad \text{if } D^i(s-1) \geq 0 \text{ or}$$

$$R^i(s) + Ar^i(s) - A^i(s) = \{Ar^i(s-1) + Ar^i(s) - \{A^i(s-1) - R^i(s-1)\}\} - A^i_{free}(s) = 0$$

$$\text{if} D^i(s-1) < 0 \tag{4.2}$$

The optimal distributions of the free capacity in the above equations can be written as the following:

$$A^i_{free}(s) = \{Ar^i(s-1) + Ar^i(s)\} \text{ if } D^i(s-1) \geq 0 \text{ or;}$$

$$A^i_{free}(s) = \{Ar^i(s-1) + Ar^i(s) - \{A^i(s-1) - R^i(s-1)\}\} \text{ if } D^i(s-1) < 0 \tag{4.3}$$

The optimal assignments of the free capacity expressed in (4.3) are generally not attainable given the fact that the free capacity in any frame is limited. The uncertainties in the values of $Ar^i(s-1)$ and $Ar^i(s)$ in the above conditions make achieving this optimal distribution of the free capacity infeasible in every UT.

However, (4.3) provides a new perspective in characterizing the packet arrival trends in the UT's. The optimal relationship provides an implicit expression for the relationship between $A^i_{free}(s)$ and the arrival trends that will minimize the throughput loss due to inaccurate assignment in addition to minimizing the average packet delay and inter-frame jitter per frame. To express this mathematically, let $A^i_{free}(s) = k(Ar^i(s-1) + Ar^i(s))$. Then (4.2) becomes:

$$R^i(s) + Ar^i(s) - A^i(s) = (Ar^i(s-1) + Ar^i(s))(1-k) \text{ if } D^i(s-1) \geq 0 \text{ or;}$$

$$R^i(s) + Ar^i(s) - A^i(s) = (Ar^i(s-1) + Ar^i(s))(1-k) - (\{A^i(s-1) - R^i(s-1)\}) \text{ if } D^i(s-1) < 0$$

If $k = 1$, then the optimum condition is obtained, i.e., $\{R^i(s) + Ar^i(s) - A^i(s)\} = 0$. This gives optimum performance in that it results in optimum channel utilization or throughput with minimum packet delay and jitter.

If $k < 1$, then $R^i(s) + Ar^i(s) - A^i(s) \geq 0$ if $D^i(s-1) \geq 0$ or;

$$R^i(s) + Ar^i(s) - A^i(s) \geq 0 \text{ if } (Ar^i(s-1) + Ar^i(s))(1-k) \geq (\{A^i(s-1) - R^i(s-1)\})$$

This condition also achieves maximal channel utilization but at possible significant packet delay since it indicates insufficiency in the capacity allocation. If $k > 1$, then $R^i(s) + Ar^i(s) - A^i(s) < 0$ resulting in inefficient assignment.

The formulation above becomes more relevant considering the fact that the actual values of $Ar^i(s)$ are not available at the frame cycle s and the value of the two-frame arrival momentum $\{Ar^i(s-1) + Ar^i(s)\}$ has to be estimated from the sequentially corresponding past values $\{Ar^i(s-1) + Ar^i(s-2)\}$, $\{Ar^i(s-2) + Ar^i(s-3)\}$ ... . Hence, typically, all the conditions are not attainable because of a combination of the uncertainties in the estimated value of $\{Ar^i(s-1) + Ar^i(s)\}$ and the limitation in the value of $F(s)$ the free capacity available in frame s. The estimation procedure in EPR-CFDAMA uses m most recent values of the time series $\{Ar^i(s-1) + Ar^i(s-2)\}$, $\{Ar^i(s-2) + Ar^i(s-3)\}$..., to predict the two-frame traffic momentum $\{Ar^i(s-1) + Ar^i(s)\}$. The algorithmic description of the EPR-CFDAMA prediction is given in the following:

Let $Z^i(s) = \{Ar^i(s-1) + Ar^i(s)\}$ and $Z^i_{est}(s) = \{Ar^i(s-1) + Ar^i(s)\}_{est}$

$$Z^i_{Peak}(s) = MAX(\{Ar^i(s-1) + Ar^i(s-2)\}, \{Ar^i(s-2) + Ar^i(s-3)\}, ...\{Ar^i(s-m) + Ar^i(s-m-1)\})$$

$Z^i_{Peak}(s)$ represents the peak two-frame arrival within the m most recent observations of the time series.

Let $\Delta Z^i(s-1) = Z^i(s-1) - Z^i(s-2) = \{Ar^i(s-1) - Ar^i(s-3)\}$

$$Z^i_{est}(s) = Z^i(s-1) + \frac{Z^i(s-1)}{Z^i_{Peak}(s)} \Delta Z^i(s-1) \quad Z^i_{est}(s) = 0 \text{ if } Z^i_{est}(s) < 0 \tag{4.4}$$

The UT weights $w^i(s)$ and free assignments $A^i_{free}(s)$ are calculated from the estimated traffic momentum according to the following expressions:

$$M^i(s) = D^i(s-1) + Z^i_{est}(s)$$

$$w^i(s) = \frac{M^i(s)}{\sum_{i=1}^{n} M^i(s)}$$

$$A^i_{free}(s) = w^i(s)F(s) = \frac{M^i(s)}{\sum_{i=1}^{n} M^i(s)} F(s) \tag{4.5}$$

$A^i(s) = D^i(s-1) + A^i_{free}(s)$

The traffic trend prediction according to (4.4) is based on the packet generation momentum that is identifiable from statistically significant events such as peak or average packet sizes, peak or average inter-arrival times over two-frame periods. These statistical properties contribute to the cumulative packet arrivals observed per frame.

In (4.4), the peak two-frame cumulative arrival property was employed. The ratio $\frac{Z^i(s-1)}{Z^i_{Peak}(s)}$ determines the existence of a statistically significant event in the two-frame period. It measures the relative value of the most recent observation in the time series to the peak value of the observations in the m most recent values. This can also be described as a measure of activity in the source. The first-order difference $\Delta Z^i(s-1)$ measures the increment trend in the two-frame cumulative arrival. A negative value of $\Delta Z^i(s-1)$ indicates a decrease in the packet generation activity over a two-frame period which is weighted by $\frac{Z^i(s-1)}{Z^i_{Peak}(s)}$. A positive value indicates a recent increase in traffic activity over a two-frame period which is also weighted by $\frac{Z^i(s-1)}{Z^i_{Peak}(s)}$. The overall momentum expression combines the residual size of a UT queue and the estimated value of the two-frame traffic increase.

4.4 Simulation and performance discussions

Using the same simulation platform that was described in chapter three, we present simulation results to show the trend prediction effectiveness of the EPR-CFDAMA protocol and the effect of this on its overall packet delay and delay jitter performance. We utilize the two types of traffic models employed in the simulations discussions in Chapter 3 and a similar TDMA/MF-TDMA return channel on the GEO BSA system described therein. The packet format for the control packets and the PDU format are also the same as described in chapter 3.

4.4.1 UT-to-GES average packet delay (ON-OFF model)

Figure 4.1 below shows the average UT-to-GES packet delay for the simulation.

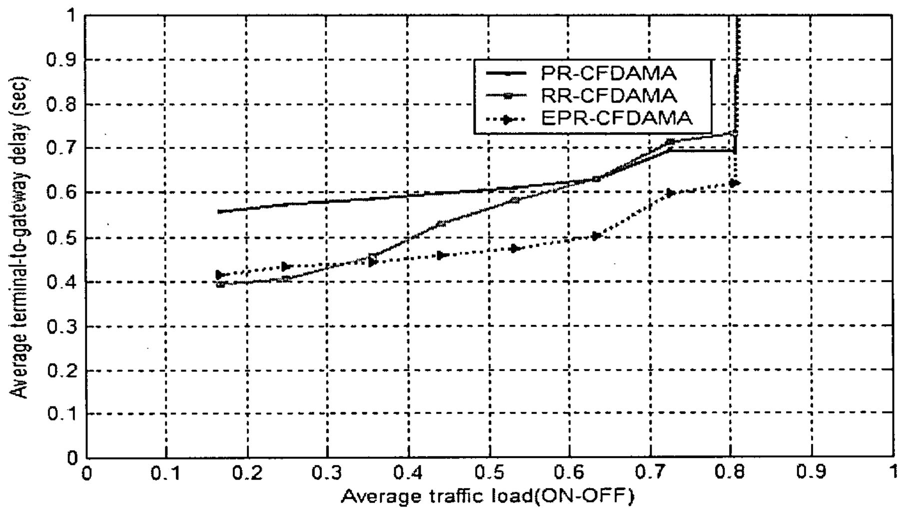

Figure 4.1: Average UT-GES packet delay as a function of average traffic load (ON-OFF)

The EPR-CFDAMA protocol provides an overall average packet delay performance that is superior to both the PR-CFDAMA and the RR-CFDAMA protocols. The superiority of the EPR-CFDAMA protocol performance, which is significantly over a broad range of traffic loads, can be attributed to the two-frame traffic momentum prediction method employed in the EPR-CFDAMA protocol resulting in free capacity assignments that takes into consideration future arrivals in the UT's.

4.4.2 UT-to-GES average packet delay jitter (ON-OFF model)

Figure 4.2 below displays the average UT-to-gateway average packet delay jitter. A performance improvement trend similar to the average packet delay is observed for the EPR-CFDAMA protocol. The superior delay jitter performance of the protocol can again be attributed to the two-frame traffic momentum prediction algorithm discussed. Since the two-frame traffic momentum algorithm considers the arrivals in two successive frames in the free capacity allocations, the high jitter-inducing effect of the inter-frame packet delay in the EPR-CFDAMA protocol can be limited thereby ensuring a better packet delay jitter performance.

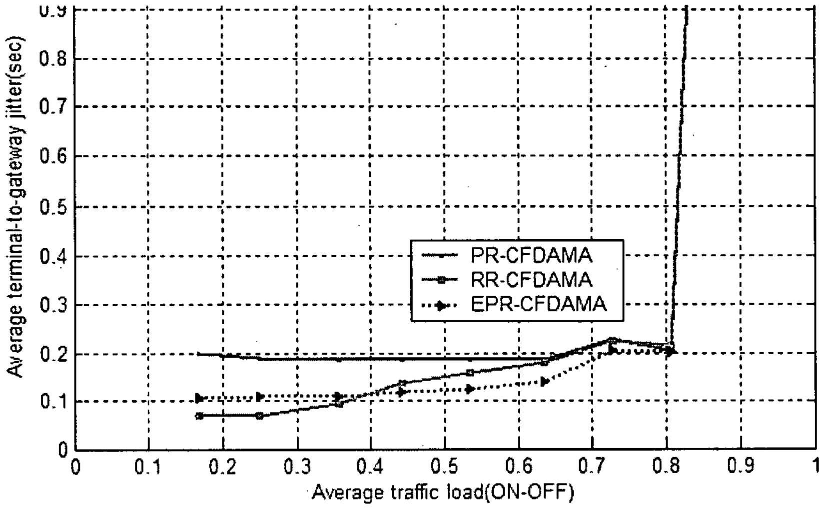

Figure 4.2: Average UT-GES packet delay jitter as a function of average traffic load (ON-OFF)

4.4.3 Statistical average performance with respect to the sufficiency parameter β and insufficiency parameter δ (ON-OFF model)

The superiority of the EPR-CFDAMA protocol over the RR-CFDAMA and PR-CFDAMA protocols as described in the delay and delay jitter performances can also be explained in terms of the sufficiency and insufficiency criteria. As shown in Figure 4.3 below, the EPR-CFDAMA protocol satisfies the capacity sufficiency criterion better than the RR-CFDAMA and the PR-CFDAMA protocols over a significantly wide range of traffic load. This implies that it is able to guarantee the highest values of parameter β, which relates to the fraction of UT's with sufficient capacity.

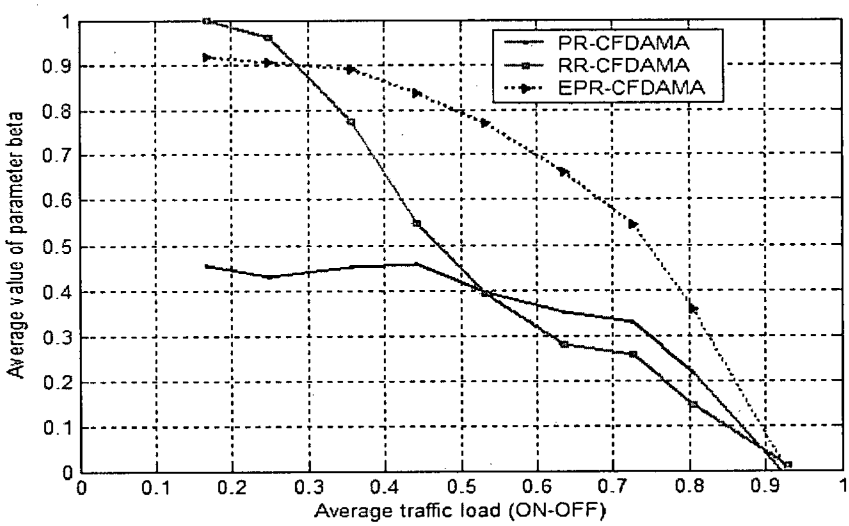

Figure 4.3: Sufficiency parameter β as a function of average traffic load (ON-OFF)

Similarly, Figure 4.4 below shows that the EPR-CFDAMA protocol can guarantee the lowest value of the insufficiency parameter δ over a wide range of traffic load. As evaluated in Chapter 3, the situation $\left(\left\{Ar^i(s-1) - A^i_{free}(s)\right\} > n_p\right)$ with $n_p = 10$ was employed in obtaining the values of δ where δ is the fraction of UT's with the above assignment situation. The specified level of capacity insufficiency is defined by n_p. By combining the characteristic displayed in parameters β and δ, a conclusion can be made that the EPR-CFDAMA protocol, while maximizing the number of UT's with sufficient capacity allocations, minimizes the number of UT's with a specified level of insufficiency when the available capacity is strictly limited.

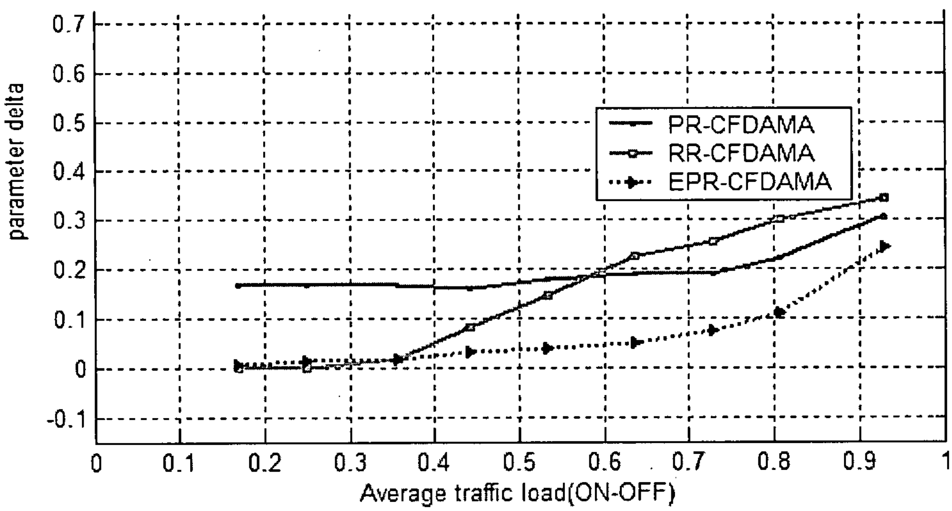

Figure 4.4: Insufficiency parameter δ as a function of average traffic load (ON-OFF)

4.4.4 UT-to-GES average packet delay and delay jitter (SFRP model)

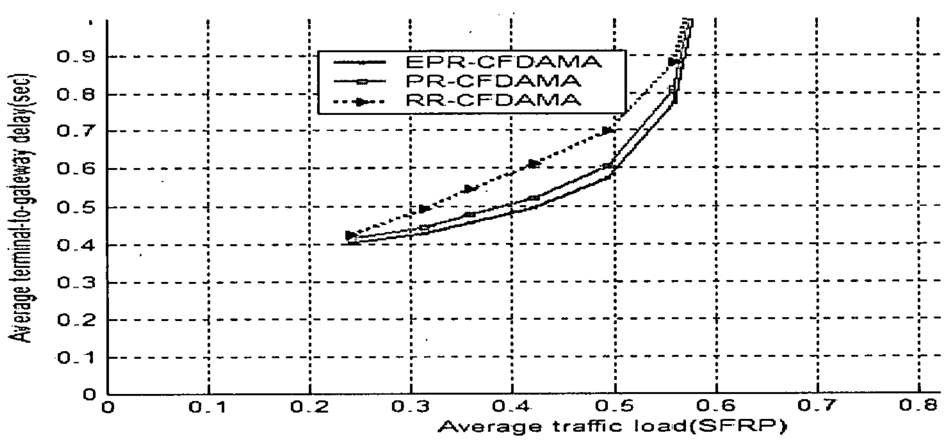

Figure 4.5: Average UT-GES packet delay as a function of average traffic load (SFRP)

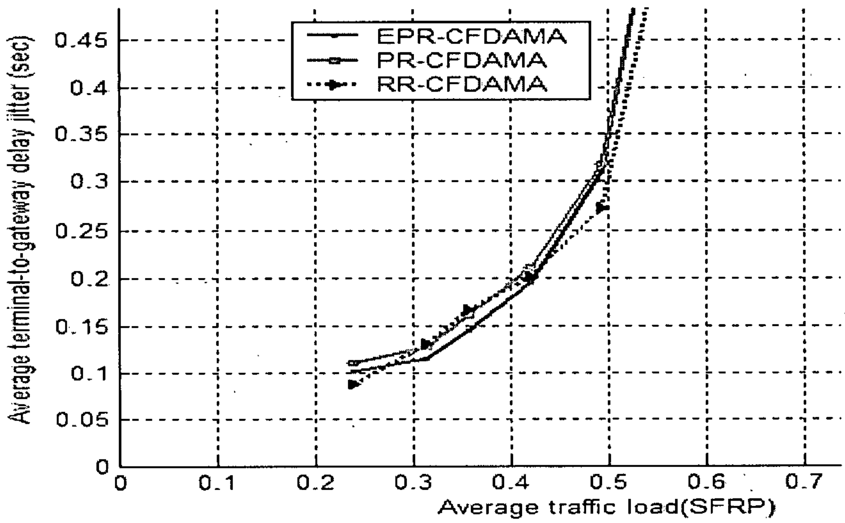

Figure 4.6: Average UT-GES packet delay jitter as a function of average traffic load (SFRP)

Figures 4.5 and 4.6, respectively, show the average packet delay and packet delay jitter performances of the protocols for the SFRP traffic model. The EPR-CFDAMA protocol's superior average packet delay performance can again be traced to the two-frame traffic momentum based prediction mechanism. Although the SFRP traffic model exhibits higher traffic variability, the prediction mechanism which is based on the two-frame arrival dynamics of the traffic source is able to better predict the trends in the UT and allocate more capacity to UT's that have higher two-frame arrival trends. This is reflected in the superior delay performance of the EPR-CFDAMA protocol when compared with the PR-CFDAMA and RR-CFDAMA protocols.

However, as shown in Figure 4.6, the delay jitter performance comparisons among the protocols can be carried out over the low-to-medium and medium-to-high ranges of traffic loads. At the low-to-medium range of traffic load, although there is significant amount of free capacity which might results in $D^i(s) = (R^i(s) - A^i(s)) \leq 0$ which tends to result in inaccurate predictions, the fact that the SFRP exhibits very high variability coupled with two-frame arrival dynamics employed in the traffic prediction enables reasonable trend predictions in the UT's, which on the overall, results in the more accurate capacity distributions to the UT's. At the medium-to-high range of traffic loads, however, the limited free capacity in the frames and the high variability in the sources results in a degradation of the delay jitter performance of the EPR-CFDAMA protocol compared with the RR-CFDAMA protocol.

4.4.5 Statistical average performance with respect to the sufficiency parameter β and insufficiency parameter δ (SFRP model)

Figures 4.7 and 4.8 show that both the PR-CFDAMA and the EPR-CFDAMA protocols can guarantee desirable values of the sufficiency parameter β and insufficiency parameter δ. They provide evidence of the correlation between the superiority of the capacity distribution efficiency of the EPR-CFDAMA protocol and its significantly better delay and delay jitter performances. Figure 4.8 shows that the insufficiency parameter of the RR-CFDAMA protocol rises sharply as the traffic load increases, a situation that results from the lack of traffic trend prediction in the RR-CFDAMA protocol.

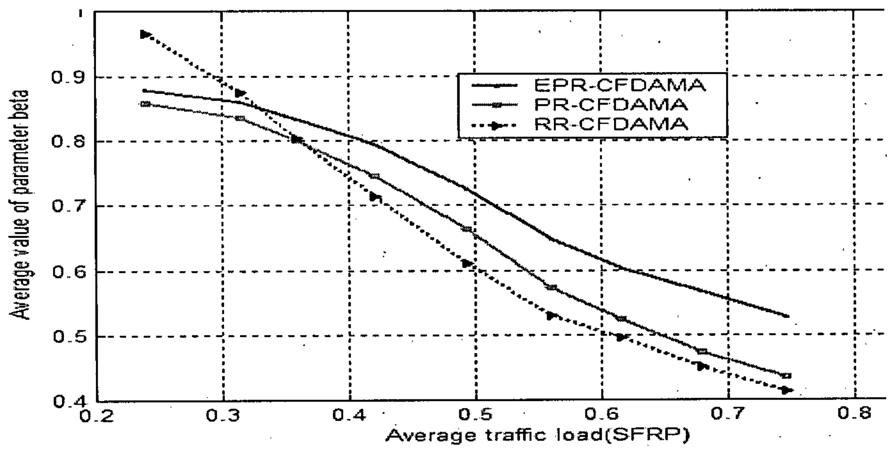

Figure 4.7: Sufficiency parameter β as a function of average traffic load (SFRP)

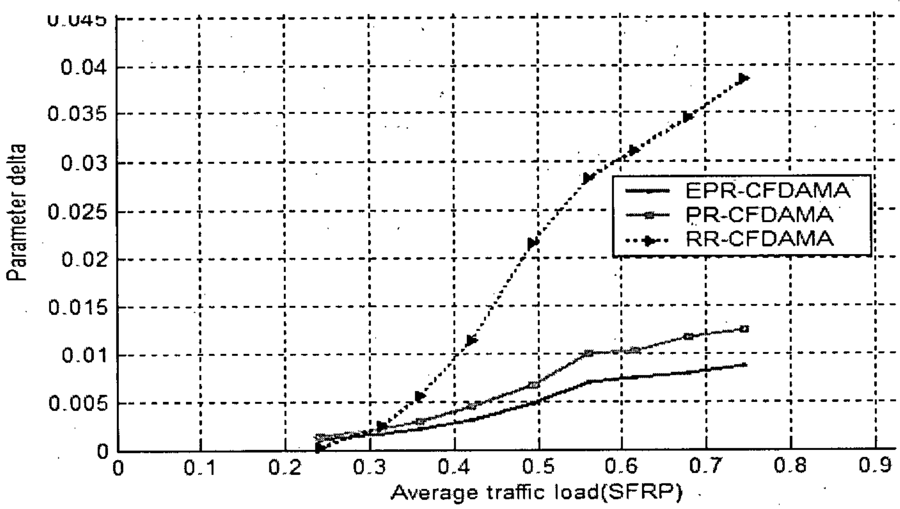

Figure 4.8: Insufficiency parameter δ as a function of average traffic load (SFRP)

4.5 Summary

The EPR-CFDAMA protocol has been analyzed using the empirical framework developed in Chapter 3. The effectiveness of the predictive mechanism of the EPR-CFDAMA protocol arises from the fact that it uses a prediction technique which is based on a time series of the two-frame packet arrivals as compared with the PR-CFDAMA protocol which uses a time series of the joint arrival-assignment dynamics of the protocol in trend prediction over the *SRTT* period. The simulation results comparing the EPR-CFDAMA, PR-CFDAMA and RR-CFDAMA protocols have been presented.

For the two types of traffic model employed, the results show that the EPR-CFDAMA protocol provides significantly better trend prediction capability over the PR-CFDAMA protocol which is evident in the significantly superior overall average packet delay and delay jitter performances. Furthermore, the results show that although the EPR-CFDAMA protocol delay and delay-jitter performances is superior to those of the RR-CFDAMA protocol for the ON-OFF traffic model, the high variability of the SFRP traffic model can significantly affect the performance of the EPR-CFDAMA protocols, especially at high traffic loads.

Given the growing trend in the use of BSA systems to support integrated or multi-class applications, it is essential for the CFDAMA protocol to appropriately optimize the performance results for the different application types in the UT's while also maximizing the overall capacity utilization of the system's capacity. Since different classes of applications have different traffic characteristics which are often related to the perceived QoS factors in the end-user applications, a potential means to achieve these goals for the CFDAMA protocol is to use application-specific traffic information in the algorithms CFDAMA protocols. To achieve this objective, we propose a cross-layer technique in the next chapter that requires a more intelligent RRS and DCA algorithm in the CFDAMA protocols.

Chapter 5 A cross-layer optimization of the CFDAMA protocols

5.0 Introduction

In this chapter, we present our analyses of the optimization of the variants of the CFDAMA protocol earlier discussed. The objective of the optimization procedure is to ensure that the CFDAMA protocol achieves differentiated MAC-layer QoS provisioning in order to be effective in a BSA system designed for multi-type and/or integrated application traffic at the UT's or GES's.

Our optimization strategy therefore utilizes application-related implementations of the two important functions performed by the CFDAMA protocol which are: RRS and DCA. Furthermore, a priority-based local capacity allocation scheme (LCAS) is applied to implement a local packet transmission schedule that follows certain observed application traffic information. Hence, the optimization strategy can be referred to as a cross-layer type.

In the following analyses, traffic differentiation and integration are two important procedures that were applied in implementing the optimization strategy. The traffic differentiation notion allows the classification of the local UT packet traffic into sets of priority-based classes or groups. The traffic aggregation notion facilitates the use of methods that allow information about packet traffic of the same type to be aggregated into the same class in the protocol algorithm. Using the empirical framework that we developed in chapter three and four, a general description of the techniques that make up the optimization of the CFDAMA protocol is presented. Subsequently, the full details of the structure of the resulting optimized protocol for a typical integrated/multi-class application structure consisting of real-time (RT) and non real-time (NRT) applications are given with performance results of the protocols obtained via simulations.

5.1 RRS in the optimized CFDAMA protocol

Figure 5.1 shows the schematic diagram of a typical UT supporting multiple traffic types or applications. We use the same frame time formulation depicted in Figure 3.1. The RRS algorithm of the Opt-CFDAMA protocol is a traffic-class-based algorithm that can utilize both the traffic differentiation and aggregation principles discussed in Chapter 2.

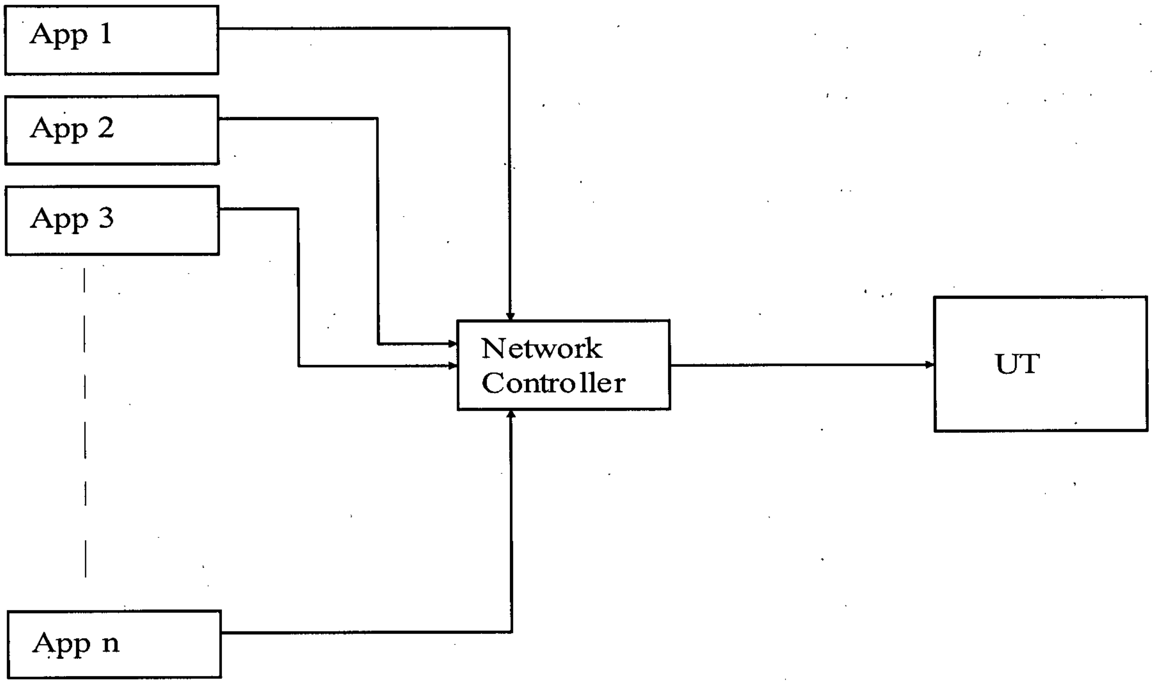

Figure 5.1: Schematic diagram of a multi-type application system in UT's

The packet traffic from all the sources in the UT are classified according to a defined system-wide traffic classification scheme. Traffic differentiation in the Opt-CFDAMA RRS algorithm is thus implemented in the UT's, each of which observes the traffic from the different sources and classifies the traffic sources according to the defined classification scheme. The classification scheme can be based on the observations of packet-level protocol information obtainable from the application-layer or transport-layer protocols employed in the end-user application/transport protocol architectures. Traffic aggregation is implemented by placing packets from different sources but defined to be of a similar traffic type (hence, of the same class) into sub-queues of the same class mainly for the purpose of making the eventual capacity requests. The Opt-CFDAMA protocol can use the same control packet format described in Chapter 3.

For the RRS, each class has its associated sub-queue size $R_j^i(s)$, i.e., size of sub-queue (class) j in terminal i at the beginning of frame s and packet arrivals from sources defined to be of the same class are accumulated into the total arrivals for each class in each UT according to the following equations:

$$R^i(s) = \sum_{j=1}^{n_{tc}} R_j^i(s)$$

$$Ar^i(s) = \sum_{j=1}^{n_{tc}} Ar_j^i(s)$$

where n_{tc} = number of traffic classes chosen for the protocol. Thus the demand in frame $(s-1)$ is earlier defined as: $D^i(s-1) = R^i(s-1) - A^i(s-1)$.

The traffic differentiation principle is implemented using a traffic table at the UT that describes the various applications observed in each UT.

Table 5.1: A proposed traffic table describing application sources in the UT's

	UT-1	UT-2	....	.....	......	.....	....	......	UT-n
UT ID	1	2	......	......	......	......	....	......	n
CC			......	......	......	......	....	......	
CID			......	......	......	......	....	......	
CR									
CAC/ APID/ AINFO									

Table 5.1 presents a possible implementation of the fields in the traffic table:

UT profile

UT ID: a unique identification number for the UT;

Class count (CC): number of classes. Each UT must categorize its local applications into one or more of the a number of system-wide predefined application classes;

Class profiles

Class ID (CID): a unique identification number assigned to a class;

Class request (CR): residual size of sub-queues of a specific class;

Class application count (CAC): number of applications in a class;

Application profiles/Information (APINFO)

Application ID (APID): a unique sequentially assigned identification number given to an application in a class;

Application information: contains the cross-layer information gathered for each application in each class.

The exact nature of the traffic information provided in the application information field is left undefined to allow for possible variations in the implementation. The definition of the classes in the application class-profile field is also not specified. Later in this chapter, we will provide a typical implementation of Table 5.1 to demonstrate how it can be used for the optimization of an integrated RT/NRT application traffic structure in the UT's. Each UT must continuously monitor its local traffic and update the class and application profile information in the table as the observed type (or class) and number of applications change.

5.2 DCA in the Opt-CFDAMA protocol

The discussions and analyses of the variants of the existing CFDAMA protocols (RR-CFDAMA, W-CFDAMA, PR-CFDAMA and EPR-CFDAMA) in the literature provide no specific descriptions on the suitability of the CFDAMA protocol as a MAC protocol to address the QoS requirements of multi-service/integrated application UT's in a BSN network. The capacity distribution efficiencies of the various CFDAMA protocols have been evaluated in terms of their overall packet delay and packet delay jitter performances. At this juncture, there is no motivation for further evaluation of the W-CFDAMA protocol given that its performance has been shown in the past two chapters to be the worst among the CFDAMA protocol variations.

Furthermore, the proposed optimization strategy should ultimately be implementable with minimal overhead when compared with the existing implementations of the various protocols. The two fundamental features of the cross-layer design in the optimized DCA algorithm are:

i. Differentiated capacity allocation: the UT's in the system receive different proportion of the uplink channel capacity according to the variations in the observed characteristics in the UT traffic: in each UT, the observed characteristics vary with variations in the behavior of the component applications;

ii. Integrated traffic prediction: the traffic trends in the UT's are to be predicted based on the overall aggregate traffic.

These two features appear contradictory because a differentiated capacity allocation based on the variations in the component applications require that traffic predictions be based on variations in the individual traffic of each application. Thus, the major challenge in achieving the cross-layer optimized CFDAMA protocol that can be readily implemented in current BSA systems is that of achieving a DCA that is sensitive to the dynamics of each application in each UT but as well, make system overhead and complexity minimal.

A realization of the afore-mentioned goal lies in recognizing that the UT's are better informed of the various applications they support and can reasonably enhance the sensitivity of the DCA to the variations in the component applications through the use of the traffic information such as the presented in Table 5.1. Thus, even though the traffic prediction is implemented based on an integrated request, each UT can effectively perform a local re-distribution of the assigned capacity to meet the requirements of the component applications. In effect, the Opt-CFDAMA protocols can utilize the same DCA algorithms already in existence for each variant. The distinguishing features of the Opt-CFDAMA protocol thus lie in an intelligent buffer management or local capacity redistribution algorithm that facilitates the transmission scheduling of the packet traffic from the different applications to utilize the assigned capacity in each UT.

5.3 Optimization with the priority-based LCAS in the Opt-CFDAMA protocol

The priority-based LCAS implements a buffer management service to re-distribute the assigned capacity to the various application traffic sources in a UT to efficiently utilize the capacity and effectively control the diverse QoS requirements of the constituent applications in a UT. The algorithm uses some or all of the information provided in Table 5.1 to implement traffic prioritization and isolation.

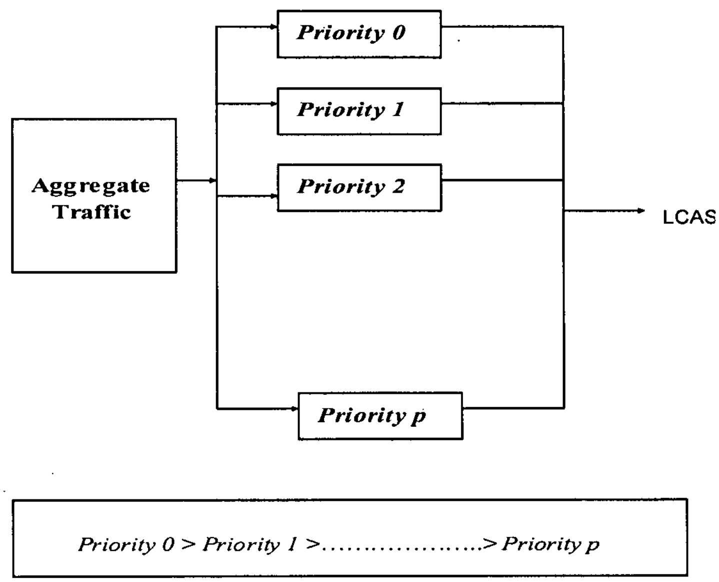

Figure 5.2: Traffic prioritization and isolation in the LCAS

Figure 5.2 shows the implementation of an integrated traffic prioritization and isolation strategy in a UT. The goal of the LCAS is to isolate each class (as identified in Table 5.1) from another via class prioritization. The class profile information readily identifies the priority levels assigned arriving packets from the traffic sources and the sub-queues to which the packets will be placed. Such information can be obtained by simple packet header checking.

The application profile information provides details specific to each application in each class. Typical application profile information required by the LCAS includes packet sizes and packet time-stamps that indicate the time that individual packet were created by the application source. The packet-level information obtained for each application in each class can then be used to locally distribute the assigned capacity and schedule the traffic from each application in order to achieve the goals of a distributed QoS for the different application types.

Thus, the LCAS can also contribute to the overall enhancement of the average packet delay or delay-jitter performance of the Opt-CFDAMA protocol. Given that in frame s the capacity $A^i(s)$ has been assigned to a UT, then the assignment can be redistributed as shown in Figure 5.2. Each class assignment $A^i_j(s)$ is based on the priority assigned to the class. If there are p numbers of priority levels in the protocol system then the following holds: $p = n_{tc}$. As shown in Figure 5.2, a higher priority class is assigned capacity first based on the current sub-queue size in the class. Each one-step-lower priority class is assigned the remaining capacity subsequently. This implicitly provides the class isolation strategy since higher-priority traffic can be assured of assignment before lower-priority traffic and lower-priority traffic only receive part capacity $A^i(s)$ if there is no high-priority packet traffic.

To ensure that the capacity assigned to all the class sub-queues are efficiently used, the class assignments $A^i_j(s)$ can be dynamically re-distributed. Within each class, the various traffic sources have access to the class capacity. Access to class capacity is controlled by a procedure that ensures that packets with the minimum timestamps are guaranteed transmission. Furthermore, any class can utilize the excess capacity assigned to another class regardless of the priority levels. To explain this analytically, the following is presented.

In the transmission period of a UT, two factors determine the transmission eligibility of a packet from a given application traffic source – its timestamp relative to the other sources in its class and the availability of capacity for the sources in its class. For a packet in class j, with timestamp t_i, the packet is transmitted in the transmission slots of the UT if and only if there are no higher-priority packets and if and only if t_i is the minimum in class j. If more than one packet from different application sources has the same timestamp, then the packet with the maximum packet size is considered eligible for transmission.

Finally, if the capacity of class j is not sufficient to guarantee the transmission of the eligible packet, the LCAS schedules the packet with some of the excess capacity of any other class regardless of it priority. The LCAS always looks for excess capacity in lower priority classes before higher-priority classes. The excess capacity typically exists if at the instant the eligibility of the packet for transmission is determined, there exists at least one class sub-queue with no packet but with assigned capacity.

5.4 The RTP/RTCP

The RTP provides end-to-end delivery services for data with real-time characteristics such as interactive audio and video. The important services provided by the RTP protocol include payload identification, sequence numbering, time-stamping and delivery monitoring. Real-time applications employing the RTP protocol typically run on top of the UDP to make use of its multiplexing and checksum services.

To monitor the QoS of packet transmission and reception and to convey information about the end-users in an RTP session, RTP has an accompanying protocol, the real-time control protocol (RTCP). RTCP is based on the periodic transmission of control packets to all the end-users in a RTP session, using the same distribution medium as the RTP data packets. The two RTP and RTCP packets are typically multiplexed in the medium by employing separate ports numbers with the UDP [11]. Among other actions, RTCP performs the following important functions:

i. The primary function is to provide feedback on the quality of the data distribution. This is an integral part of RTP's role as a transport protocol and it is related to flow and congestion control functions of other transport protocols. The feedback may be useful for adaptive encoding control [11] or to diagnose faults in the distribution;

ii. RTCP carries a persistent transport-level identifier for an RTP source called the canonical name or CNAME. This is required in order to keep track of each end-user in an RTP session if the SSRC identifier changes. The CNAME also enables multiple data streams (e.g., audio and video streams) from the same end-user to be associated for effective inter-media synchronization in receiving terminals.

The following are the features of the RTP protocol that makes it useful in providing cross-layer information for application-specific services in lower-layer protocols:

i. It is designed to be malleable to provide the information required by a particular application and is integrated into the application processing ;

ii. The RTP protocol can be tailored to various applications via modifications and/or additions to the headers as needed.

Thus, the RTP protocol can support a wide range of multimedia encoding standards (MPEG-2, MPEG-4, H.263 etc) that are used for video applications.

5.5 Integrating the Opt-CFDAMA protocol in an IP-based BSA system with RTP/RTCP and TCP-type applications

On the basis of MAC layer protocol performance, applications traffic over current BSA systems can be categorized into two broad types: real-time (RT) and non-real-time (NRT) traffic. In particular, multimedia applications, consisting of a combination of video and (or) voice applications transported by RTP/RTCP and data applications transported by TCP can generally be classified respectively into these two classes of traffic with varying MAC-layer QoS requirements.

The Opt-CFDAMA protocol, with its differentiated capacity allocation, integrated traffic prediction and TS capabilities can be effectively integrated with the existing IP-based service architectures (DiffServ) for multi-service applications over BSA systems. The DVB-RCS system further presents avenues for the Opt-CFDAMA protocol to be utilized in current and future satellite systems. Figure 5.3 presents a proposed service provisioning architecture with which the Opt-CFDAMA protocol can be integrated in an IP-based BSA system. The DVB-RCS system proposes a dynamic capacity assignment in the following categories: continuous rate assignment (CRA), rate-based dynamic capacity (RBDC), volume-based dynamic capacity (VBDC), absolute volume-based dynamic capacity (AVBDC) and free capacity assignment (FCA) [3],[10]. In the Opt-CFDAMA protocols, the CRA, RBDC, VBDC and AVBDC can all be used in the RRS to schedule the initial demands of each UT. The free capacity assignment methods of each of the variants of the Opt-CFDAMA protocols can then be used in a FCA method. Thus, the Opt-CFDAMA protocol can be regarded as an enhancement to the FCA method in the DVB-RCS system.

The most important aspect of the integration of the Opt-CFDAMA protocol according to the service architecture is the implementation of the application and class profile fields described in Table 5.1. The application profile fields identify the various applications in a UT and the class profile fields identify the various classes of applications that have been defined for the system. Both fields can then be employed in the RRS algorithm and the LCAS of the Opt-CFDAMA protocols. For the proposed architecture, the applications can easily be identified as either belonging to an RT-type or NRT-type. To establish the type of an application and define the classes of applications in a UT, the proposed scheme can use the protocol information obtainable from the IP protocol and the DiffServ domain. DiffServ is an IP-based service protocol that support differentiated service provisioning through the use of a number of per-hop-behaviors (PHB) which identify the service handling procedures for packet flows in traffic. The established PHBs are as defined below:

i. Expedited forwarding (EF) PHB which typically provide low-loss, low-delay and low-jitter services;

ii. Assured forwarding (AF) PHB which requires a capacity guarantee but has no delay or jitter constraints;

iii. Default (DE) PHB which is the same as the best effort service with no requirements of any type.

These PHBs can be defined in the DiffServ domain via packet-header protocol information exchanges between the DiffServ procedure and the IP procedure.

The different PHBs can be used in the Opt-CFDAMA protocol to define the classes for developing Table 5.1 and subsequently implementing the RRS algorithm and the LCAS. Since RT application traffic employing RTP/RTCP will definitely be associated with the EF PHB while NRT application traffic can be associated with either the AF or DE PHBs, a mapping of the PHBs to application types (RT/NRT) can easily be carried out in the Opt-CFDAMA protocol. Thus, for multimedia applications in the UT's, two classes can be defined for developing Table 5.1 and implementing the RRS and LCAS.

Moreover, the LCAS will require packet-level application information for the scheduling of packet transmissions to optimize capacity usage and satisfy the diversity in the capacity requirements of the various applications. Since the QoS requirements to be differentially optimized in the MAC-layer protocol are the delay and delay-jitter requirements, the packet-layer information required in the LCAS to schedule the various application traffic in the UT's are the packet timestamps which indicate the relative packet generation times of the applications' packet traffic and the packet sizes that indicate the proportion of the assigned UT capacity that is required to completely transmit any packet. For a RT-type application transported by RTP/RTCP, this information is readily obtainable from the RTP/RTCP application profile that can be stored locally in a UT. This thesis focuses on optimizing the performance of the CFDAMA protocol to support RT RTP/RTCP-type applications in the presence of NRT TCP-type applications.

5.5.1 RTP/RTCP Opt-CFDAMA protocols

Figure 5.3 shows the overall service features of the system after the Opt-CFDAMA protocol has been integrated in a RTP/RTCP/IP/DiffServ/DVB-RCS service system. The RT-type sources are considered to be of the RT class while the NRT-type sources are of the class NRT. In the table, the CID/CAC/APID provides a mapping of the classes in the UT to the applications in the UT. With the APID/AINFO, references to the profiles of the applications in each UT can be obtained for the class-to-application mapping. The two classes are numbered 0 and 1 with 0 representing a higher priority RT class and 1 representing a lower priority NRT class. The applications can be numbered sequentially from zero up to the number of applications in each UT.

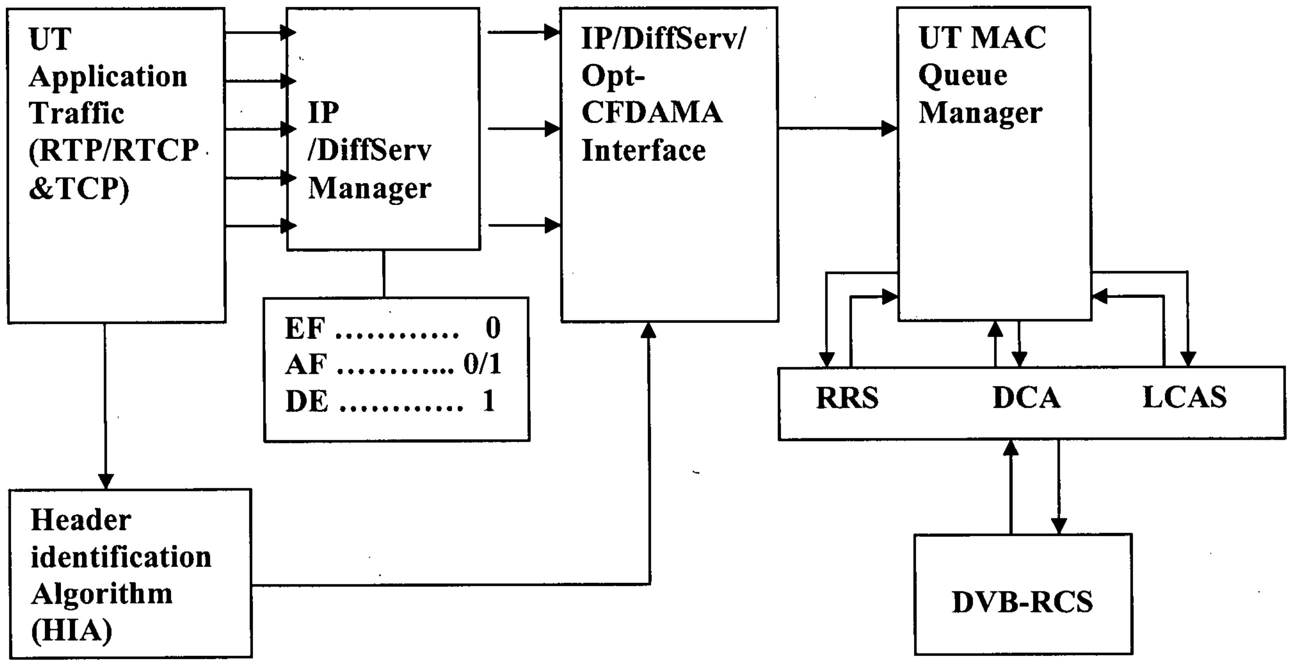

Figure 5.3: An integrated service architecture for the RTP/RTCP/TCP/IP/Opt-CFDAMA/DVB-RCS

The proposed service system is implemented in the UT processor. The IP/DiffServ/Opt-CFDAMA interface allows the PHBs to be integrated in the Opt-CFDAMA protocol for implementing Table 5.1 and the eventual classification of the traffic from the application sources as well as the eventual prioritization of the different traffic sources identified in each UT. The header identification algorithm (HIA) continuously monitors the traffic from the applications and identifies the traffic type by using locally stored protocol profiles of each type of applications. The types can then be mapped to the PHBs derived from the IP/DiffServ domain in the IP/DiffServ/Opt-CFDAMA protocol interface. RT-type traffic flows are mapped to the EF PHB and NRT-type traffic flows are mapped to the AF or DE PHBs. Each traffic source (identified from the different ports) is queued in a bank of buffers created for each class in the UT queue manager.

The FCA dynamic assignment scheme of the DVB-RCS forms the basis for the integration of the Opt-CFDAMA DCA and the DVB-RCS system. The RRS algorithm of the Opt-CFDAMA protocol obtains the resource requirements of each class in the UT manager, aggregates the requests and signals the capacity requests in the control sub-frame. The DCA algorithm (implemented in the distributed of centralized form) obtains the capacity assignments according to the Opt-CFDAMA DCA scheme, and forwards the assignment to the

UT MAC manager which implements the transmission of the packets from the different sub-queues in each class according to the packet-level protocol information obtained from each packet arriving in each sub-queue for each class. Both of these are obtainable from the RTP/RTCP protocol information profile stored locally. In our research work, we consider RTP and RTCP packets to be of the same priority, and since they are transported using different ports in UDP [11], the packets from each are placed in different sub-queues.

The LCAS uses the timestamp and packet size information to schedule the transmission of the different sources in each class according to the description in Figure 5.2. Within each class, of traffic, the timestamp information and packet size information in each packet is used to optimize the re-distribution of the assignment to minimize the overall average delay and jitter and maximize the usage of the assigned capacity.

5.6 Modeling and simulation of the cross-layer Opt-CFDAMA protocols

5.6.1 Simulation model

To evaluate the Opt-CFDAMA protocols, we employed the same network system model that was simulated in Chapters 3 and 4. However, for the Opt-CFDAMA protocol simulations, we demonstrate the existence of multi-type applications in the UT's by employing two sets of traffic sources in each UT.

5.6.2 Traffic model

We simulated three data traffic sources as the NRT traffic sources and two video sources as the RT sources. The two video models were generated from real medium quality MPEG-4 video packet traces obtained from [46]. The three NRT sources were from the same traffic models used in Chapters 3 and 4. Similar to the simulations scenarios in Chapters 3 and 4, the two types of self-similar data traffic models (ON-OFF and SFRP) were employed for traffic generation in the NRT traffic sources. Figure 5.4 below shows the schematic model of the UT's in the OPNET simulation environment. The internal working models of the sources and the UT's are similar to the ones described in Chapter 3. Similar to the procedure for increasing the traffic load in Chapters 3 and 4, the traffic load in the multi-service BSA network model was increased by increasing the number of UT's in the network. The normalized traffic load was measured for the aggregate offered traffic according to (3.12).

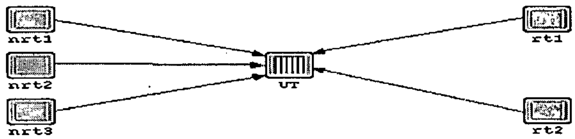

Figure 5.4: OPNET UT model configuration with three NRT sources and two RT sources

Table 5.2: Traffic Parameters of RT-1 (MPEG-4)

Frame Statistics	Unit	Value
Compression Ratio	YUV:MP4	28.40
Video Run Time	msec	3.6 e+06
Number of Frames		89998
Mean Frame Size	Byte	1.3 e+3
Variance of Frame Size		1.3 e+6
Minimum Frame Size	Byte	26
Maximum Frame Size	Byte	8511
Mean Bit Rate	Bit/Sec	2.7 e+5
Peak Bit Rate	Bit/Sec	1.7 e+6
Peak/Min Bit Rate		6.36

Table 5.3: Traffic Parameters of RT-2 (MPEG-4)

Frame Statistics	Unit	Value
Compression Ratio	YUV:MP4	97.83
Video Run Time	msec	3.5 e+06
Number of Frames		89998
Mean Frame Size	Byte	3.9 e+02
Variance of Frame Size		2.1 e+05
Minimum Frame Size	Byte	26
Maximum Frame Size	Byte	4690
Mean Bit Rate	Bit/Sec	7.8 e+04
Peak Bit Rate	Bit/Sec	9.4 e+05
Peak/Min Bit Rate		12.07

Tables 5.2 and 5.3 above show the parameters of the two video sources. We chose to measure the optimality of the protocols by evaluating the performances of the RRS, DCA and LCAS components of the optimization that we have developed for the RR-CFDAMA, PR-CFDAMA and EPR-CFDAMA protocols. The essential features of the simulation are as defined below:

i. The classification scheme for the simulation is a two-class scheme with the higher-priority class defined for the RT video sources and the lower-priority class defined for the NRT data sources;

ii. Each application traffic source is independent and each is uniquely identifiable in the UT's;

iii. The RRS algorithm performs an aggregated request based on the combination of the requests from all the classes in each UT; however, the RRS keeps a table for all the identified traffic sources;

iv. The DCA algorithm operates on the aggregated requests from each UT;

v. The priority-based LCAS uses the application and class profile information generated by the RRS algorithm to provide the local scheduling of the traffic sources as earlier described.

5.6.3 Peformance comparisons of the Opt-CFDAMA protocols

Using the above-described UT configurations, BSA network system and the proposed service architecture integrating the Opt-CFDAMA protocols with the DVB-RCS and IP/DiffServ for multimedia applications consisting of RTP/RTCP transported video applications and TCP-type data applications, we provide simulation results to characterize the distributed QoS provisioning capability of the Opt-CFDAMA protocols in an integrated application structure exemplified by multimedia applications. Our performance discussions focus on comparing the average delay and jitter for the Opt-CFDAMA protocols and cross-evaluation of the Opt-CFDAMA and the regular CFDAMA protocol variations.

The performance results to be presented show two aspects of the comparisons. Firstly, the performances of the RTP/RTCP optimized variants of the RR-CFDAMA, PR-CFDAMA and EPR-FDAMA protocols (Opt-EPR-CFDAMA, Opt-PR-CFDAMA and Opt-RR-CFDAMA) are compared. In the Opt-CFDAMA protocol versions, packets from the RT and NRT sources emanating at the UT's are placed in two classes of sub-queues one for each traffic type.

Secondly, the performances of the RTP/RTCP optimized CFDAMA protocol variations are compared against the regular versions of each CFDAMA protocol variations to show how the combination of intelligent signaling (performed by the RRS algorithm), aggregated DCA and distributed QoS provisioning (performed by the LCAS) can guarantee a differentiated service policy in the Opt-CFDAMA protocols. In this case, packets from all types of traffic sources (RT and NRT) are placed in the same queue regardless of the types of the sources. In all the simulation cases, we considered the existence of a mix of the RT-type traffic (video sources) and the NRT-type sources (data sources). Since the type to which each packet belongs can be identified, the overall average delay or delay jitter for each type of traffic sources can be readily separately obtained and compared.

5.6.3.1 Average UT-to-GES delay comparisons (ON-OFF-NRT)

Figures 5.5 and 5.6 present a comparison of the average UT-to-gateway delay for the RT-type and NRT-type applications in the Opt-CFDAMA protocols and in the regular CFDAMA protocol variations.

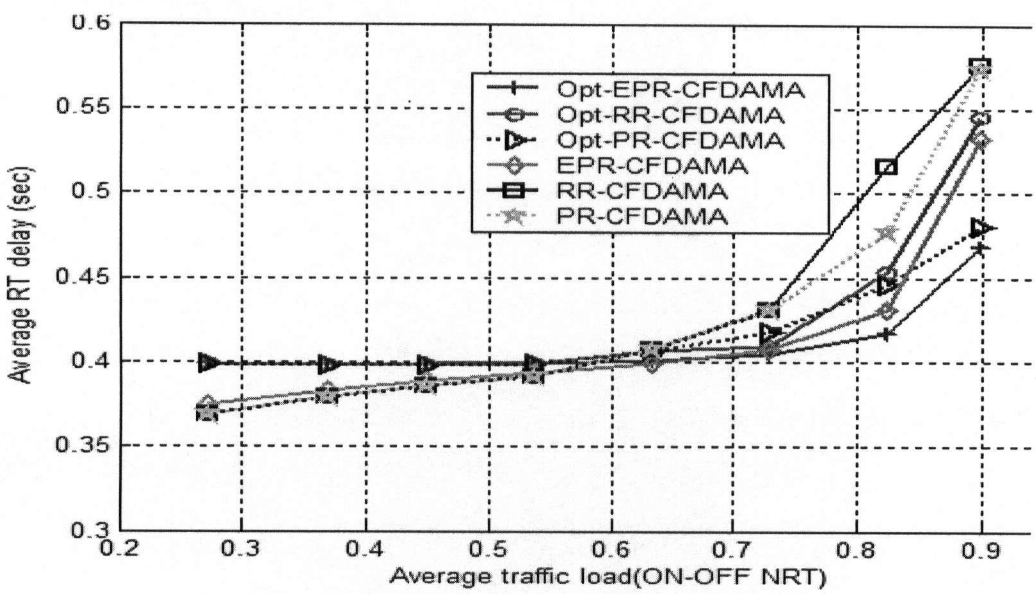

Figure 5.5: Average UT-GES RT delay as a function of average traffic load (ON-OFF NRT)

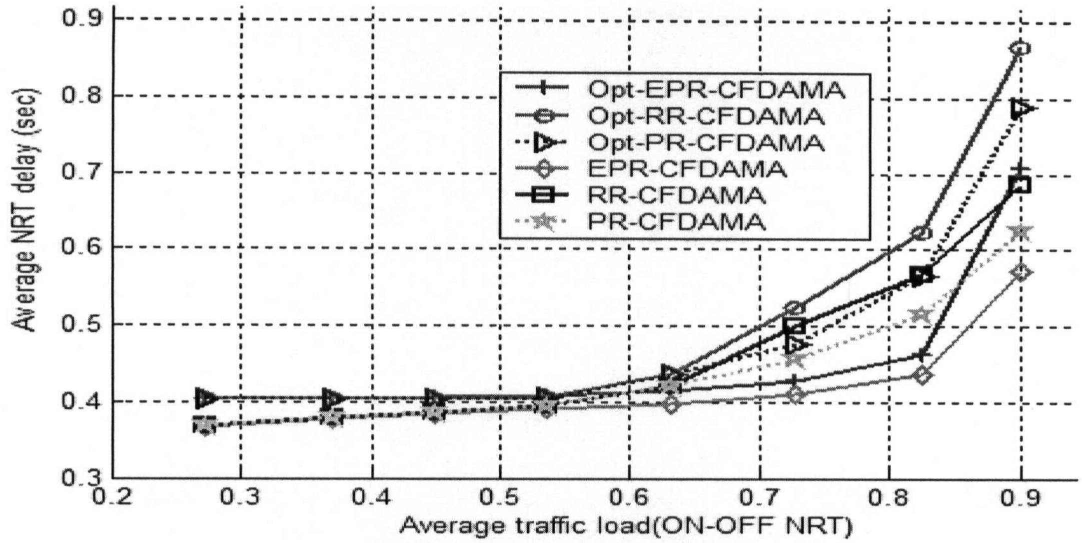

Figure 5.6: Average UT-GES NRT delay as a function of average traffic load (ON-OFF NRT)

Among the Opt-CFDAMA protocol variations and for both the RT-type and NRT-type traffic, the Opt-EPR-CFDAMA protocol shows significantly better delay performance over the Opt-PR-CFDAMA and the Opt-RR-CFDAMA. As discussed in the previous simulations in Chapter 4, the better delay performance is a result of the superiority of the prediction method of the DCA algorithm of EPR-CFDAMA. When the DCA algorithm is combined with the traffic prioritization in the LCAS in the Opt-EPR-CFDAMA protocol, at any level of traffic load, the RT-type traffic in a UT will gain more access to the assigned capacity at the expense of the NRT-type traffic. This explains why the average delay experienced by the RT-type traffic is the minimum for the Opt-EPR-CFDAMA protocol but not for the NRT-type traffic.

Moreover, the traffic prioritization in the LCAS also resulted in the NRT-type traffic experiencing higher average delays in all the Opt-CFDAMA protocol variants than in the regular CFDAMA protocol variations. The performances of the regular CFDAMA protocols all exhibit a trend consistent with the simulation results that were obtained in Chapters 3 and 4 despite the fact that the traffic model considered is a mix of the NRT-type traffic sources employed in chapters three and four and the RT-type sources consisting of the two video sources in the UT's.

5.6.3.2 Average UT-to-GES delay jitter comparisons (ON-OFF NRT)

Figures 5.7 and 5.8 present the average UT-to-gateway delay jitter performance for the Opt-CFDAMA protocols and the regular CFDAMA protocol variations.

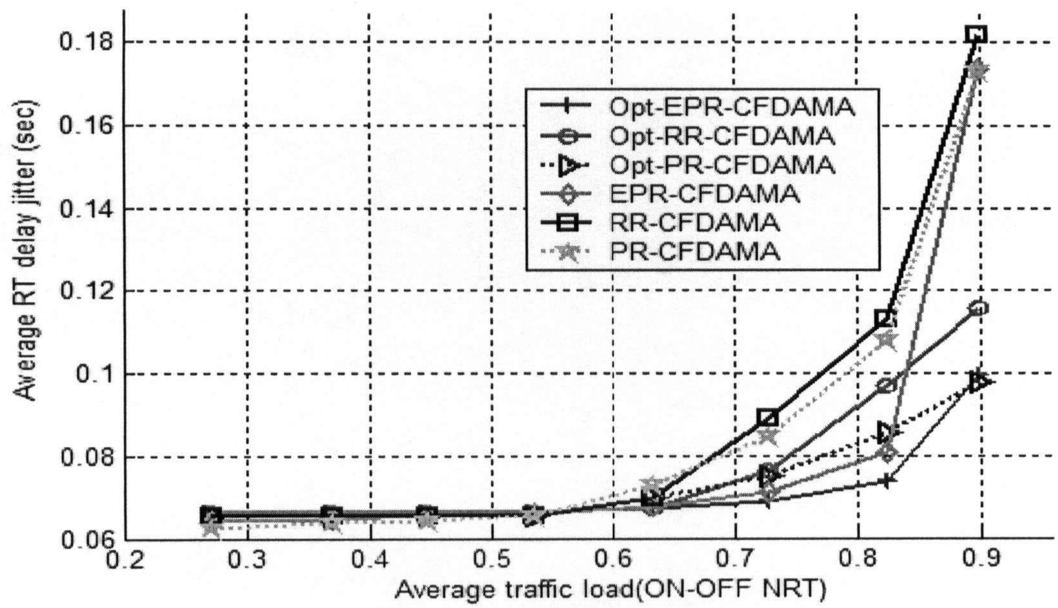

Figure 5.7: Average UT-GES RT delay jitter as a function of average traffic load (ON-OFF NRT)

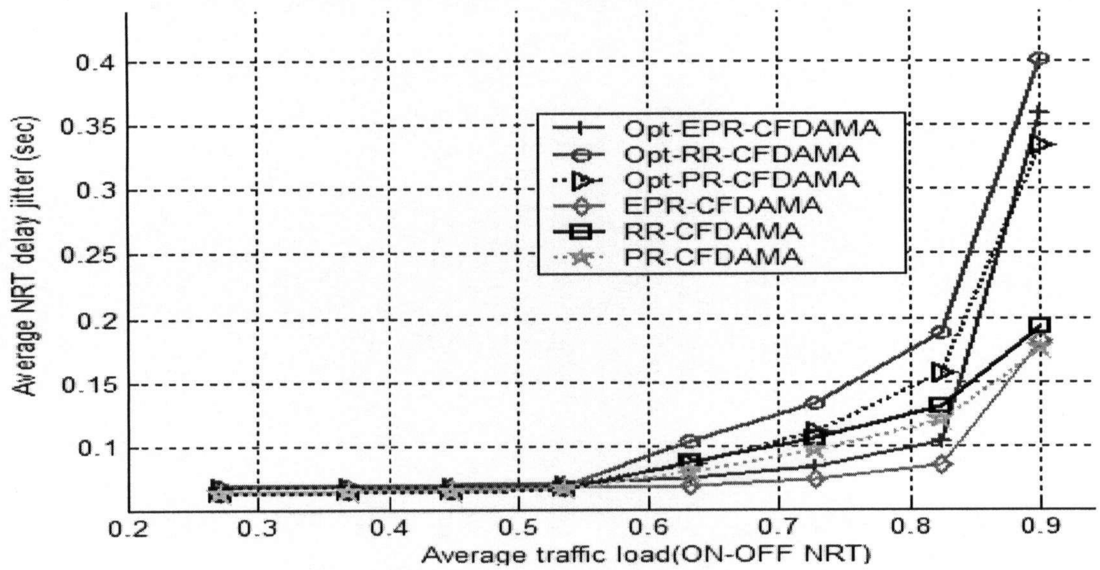

Figure 5.8: Average UT-GES NRT delay jitter as a function of average traffic load (ON-OFF NRT)

As shown in Figures 5.7 and 5.8, a performance behavior similar to that observed for the average delay is observable for the average delay jitter in the protocols for the two types of traffic in the UT's. The Opt-EPR-CFDAMA protocol is able to guarantee the minimum delay jitter performance for the RT-type applications over the low-to-high range of traffic load but with a significantly poorer delay jitter performance for the NRT-type traffic at very high loads due to the effect of the prioritization of the RT-type traffic over the NRT-type applications in the LCAS. The effect of the prioritization of RT-type traffic over NRT-type applications is noticeable in all the Opt-CFDAMA protocol variations with all the Opt-CFDAMA protocols showing significantly higher delay jitter performance for the NRT-type traffic compared with the regular CFDAMA protocol variations. The overall effect of the LCAS is to transfer the jitter of RT-type traffic onto the NRT-traffic by guaranteeing the RT-type traffic any available capacity.

5.6.3.3 Statistical average performance with respect to the sufficiency parameter β and insufficiency parameter δ (ON-OFF model)

Figures 5.9 and 5.10 characterize the network-wide capacity distribution efficiency of the CFDAMA protocols.

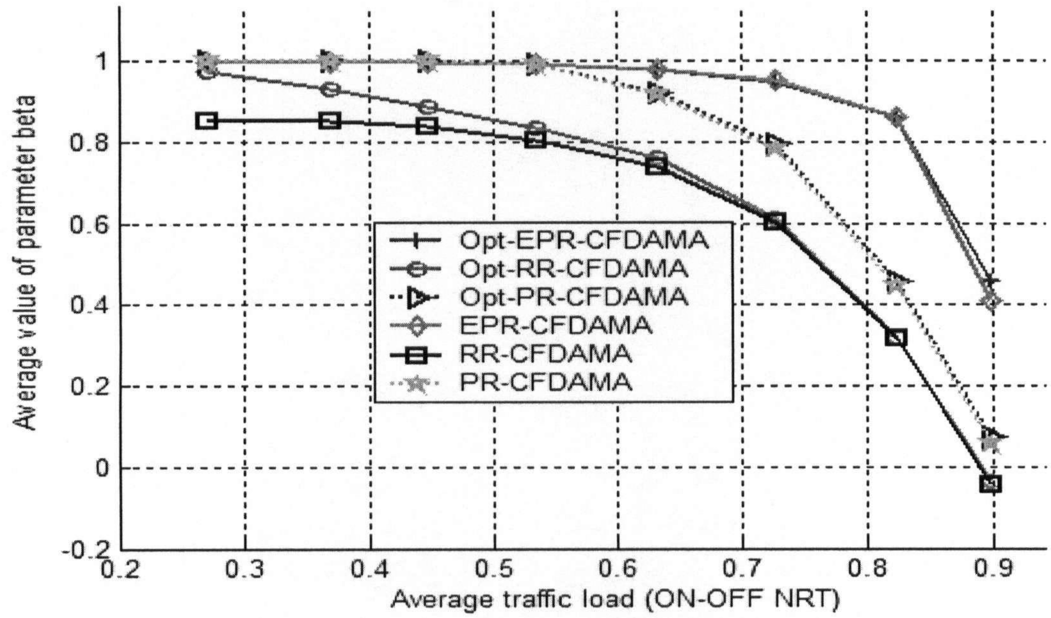

Figure 5.9: Sufficiency parameter β as a function of average traffic load (ON-OFF NRT)

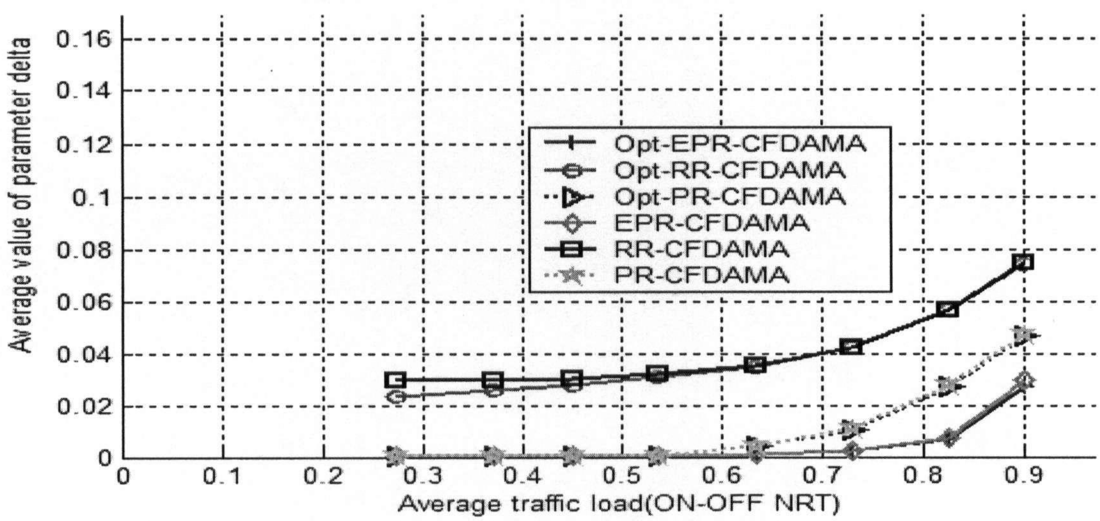

Figure 5.10: Insufficiency parameter δ as a function of average traffic load (ON-OFF NRT)

As shown in Figures 5.9 and 5.10, there are no significant distinctions between the sufficiency parameters β and insufficiency parameters δ of the Opt-CFDAMA protocols and the regular CFDAMA protocols. The proximities of these can be attributed to the fact that the DCA algorithms of both the Opt-CFDAMA protocol variants and the regular CFDAMA protocol variants are implemented on the same aggregated capacity request values.

5.6.3.4 Average UT-to-GES delay comparisons (SFRP NRT)

Figures 5.11 and 5.12 display the average delay performances of the regular and Opt-CFDAMA protocols with the NRT-type traffic modeled as a SFRP with parameters specified in Chapter 3.

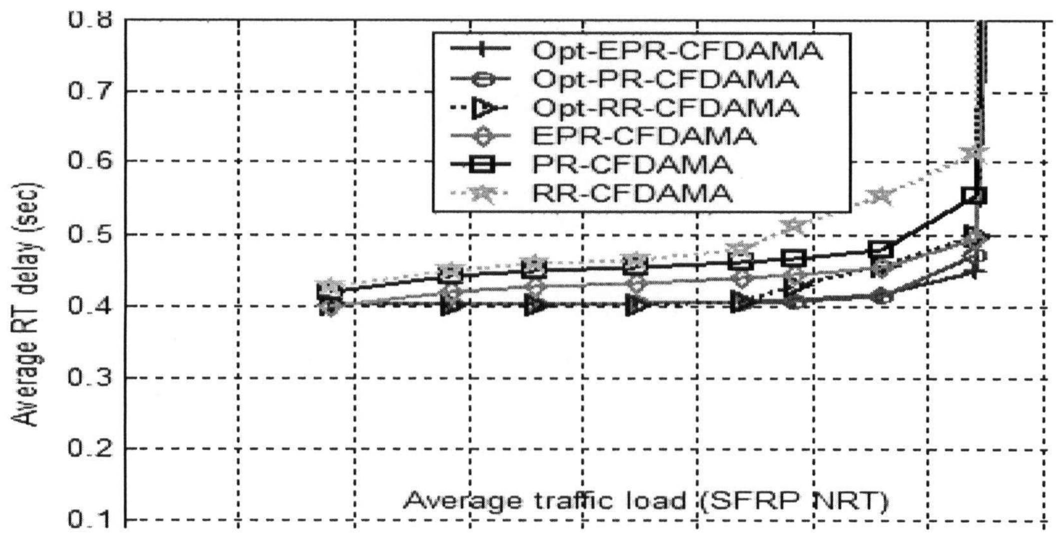

Figure 5.11: Average UT-GES RT delay as a function of average traffic load (SFRP NRT)

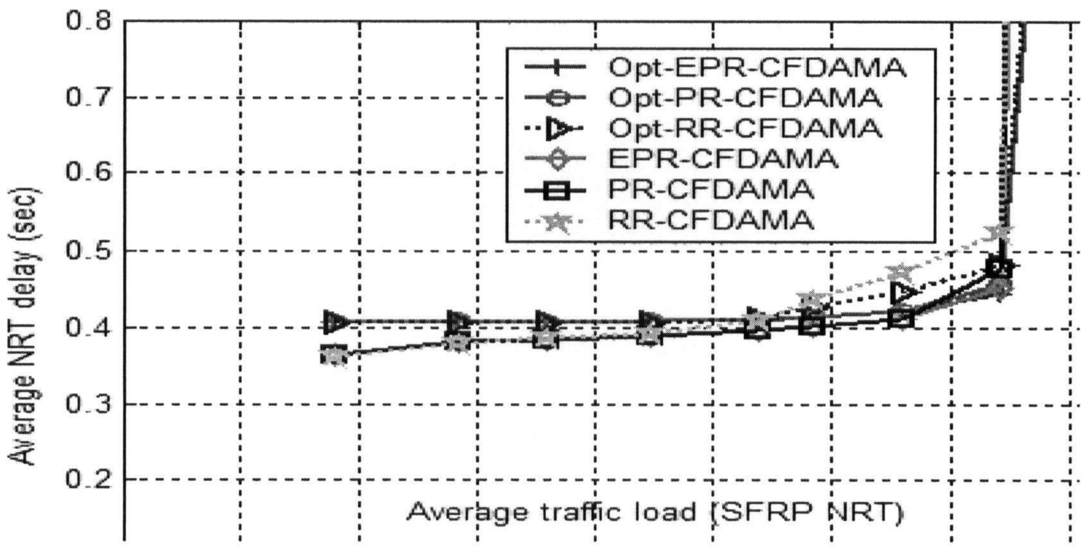

Figure 5.12: Average UT-GES NRT delay as a function of average traffic load (SFRP NRT)

The Opt-CFDAMA protocols provide a better delay performance (than the regular CFDAMA protocols) for the RT-type traffic in the simulation due to the traffic prioritization in the LCAS that guarantees more capacity access to the RT-type traffic sources. However, for the NRT-type traffic, the differences in the performances of the Opt-CFDAMA protocols are only significantly different from one another at very high traffic load. This can be attributed to the high burstiness of the SFRP traffic model. In the Opt-CFDAMA protocols with the bursty SFRP NRT traffic, the LCAS's prioritization effect on the delay performance differentiation will depend on the characteristics of the NRT-type and RT-type traffic arrivals from frame to frame. If few NRT-type and RT-type packets arrive in the UT's, the NRT-type packets are more likely to be transmitted without significant average delay differences as long as there is enough capacity available for both types to be transmitted.

In the regular CFDAMA protocols, since the packets are queued and transmitted in the order of their arrivals regardless of the traffic type, the NRT-type packets are likely to experience a similar average delay pattern as they would experience in the presence of RT-type traffic with the LCAS in effect. The only factor that will influence the delay pattern will be the available capacity which becomes significantly limited at high traffic loads.

5.6.3.5 Average UT-to-GES delay jitter comparisons (SFRP NRT)

In Figures 5.13 and 5.14, the Opt-CFDAMA protocols and the regular CFDAMA protocols exhibit performance trends similar to the ones described above. For the RT-type traffic, the delay jitter performances of the Opt-CFDAMA protocols are generally better than those of the regular CFDAMA protocols because of the prioritization and isolation effect of the LCAS.

As shown in Figure 5.13, the delay jitter performances of the Opt-CFDAMA protocols are close. This results from the prioritization effect of the LCAS in the Opt-CFDAMA protocol which guarantees the available capacity to the RT-type traffic in all the Opt-CFDAMA protocol variants in the presence of the NRT-type traffic.

In Figure 5.14, it is evident that the delay jitter performances of the Opt-CFDAMA protocols are not significantly different from their corresponding regular variants for the SFRP NRT-type traffic. As afore-explained for the delay performances, this is because of the high variability in the SFRP NRT traffic which tends to limit the prioritization effect of the LCAS in the Opt-CFDAMA protocols.

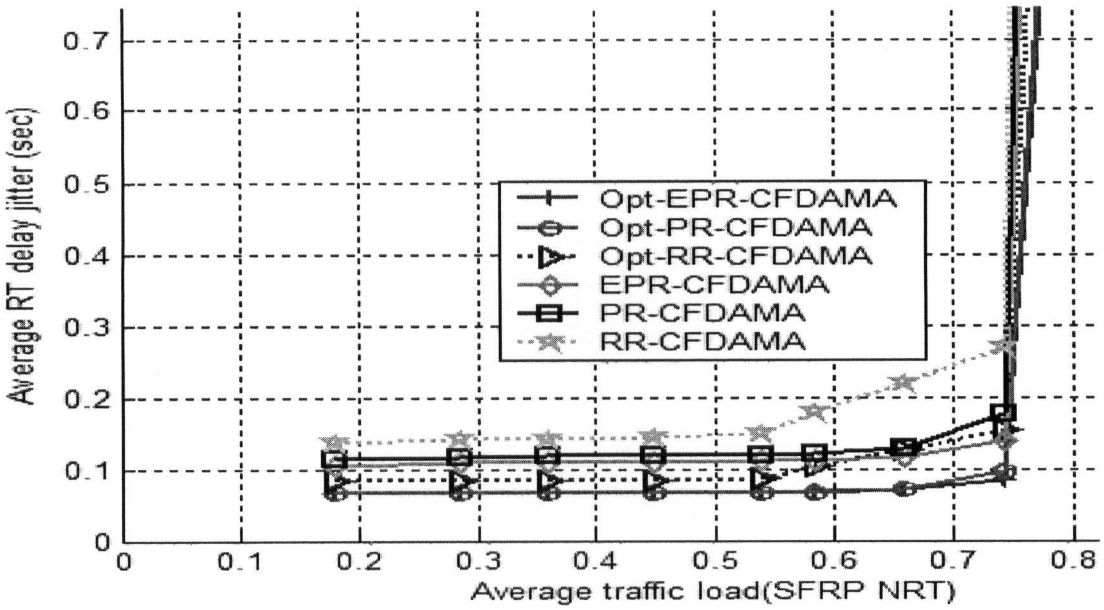

Figure 5.13: Average UT-GES RT delay jitter as a function of average traffic load (SFRP NRT)

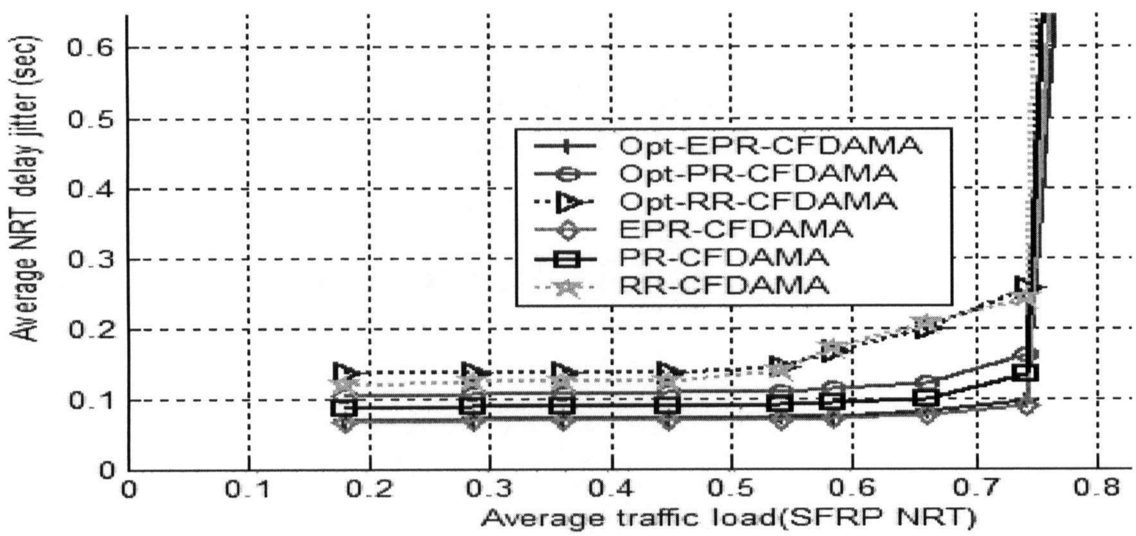

Figure 5.14: Average UT-GES NRT delay jitter as a function of average traffic load (SFRP NRT)

5.6.3.6 Statistical average performance with respect to the sufficiency parameter β and insufficiency parameter δ (SFRP model)

Figures 5.15 and 5.16 show that the sufficiency parameter β and the insufficiency parameter δ of the CFDAMA protocols and the Opt-CFDAMA protocols are not significantly different since both groups of the CFDAMA protocol utilize the same request parameter values in implementing their respective DCA algorithms.

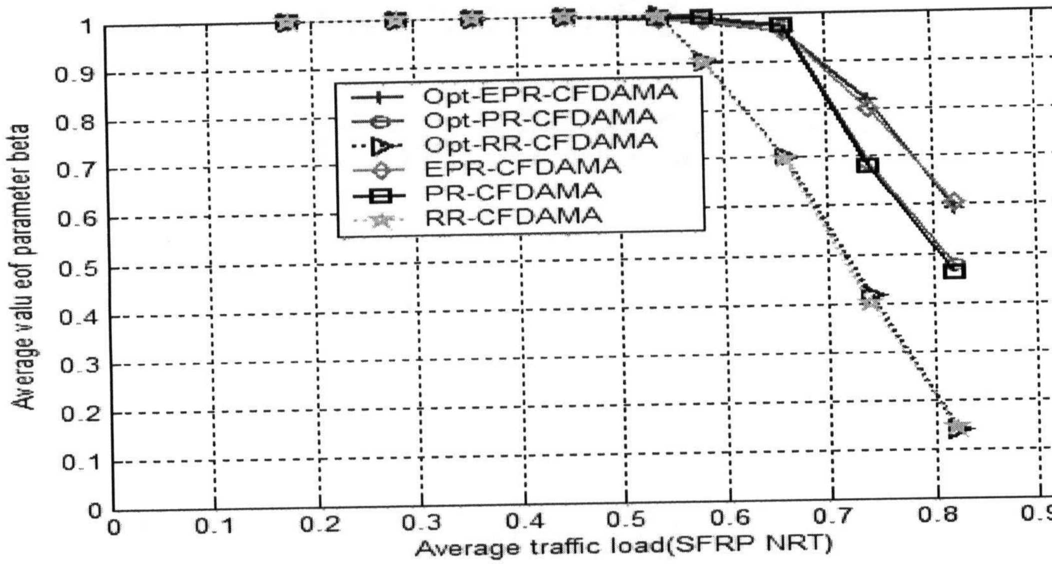

Figure 5.15: Sufficiency parameter β as a function of average traffic load (SFRP NRT)

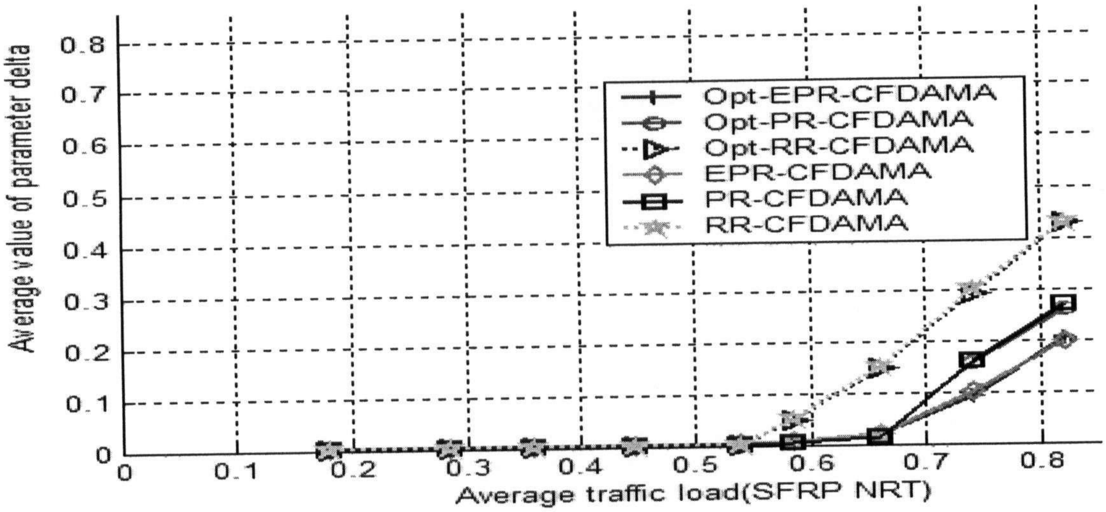

Figure 5.16: Insufficiency parameter δ as a function of average traffic load (SFRP NRT)

5.7 Summary

In this chapter, we have proposed a scheme for employing the CFDAMA protocol in a multi-type application system in UT's. The proposed scheme employs protocol information across the application layer, the IP layer and MAC layer in optimizing the performance of the CFDAMA protocol to support differential QoS provisioning for multi-type applications in the UT's. This forms an important contribution to employing the CDAMA protocol in supporting multimedia applications with varying levels of QoS requirements over the BSA infrastructure.

Additionally, to demonstrate how the optimized CFDAMA protocols can be integrated in the current IP-based BSA system, we have also proposed a service architecture which can integrate the RTP/RTCP (for RT-type applications) and the DiffServ domain in the IP for to provide traffic class identifications in the UT's while also enabling the prioritization of traffic in the UT and isolation of the RT-type application from the NRT-type application transported by the TCP. Our simulation analysis for the two-class applications system (RT/NRT) in the UT's has shown that through using traffic and protocol information to implement the RRS, and a priority-based LCAS of the CFDAMA protocols, the various MAC QoS requirements (delay and delay jitter) of the applications in a UT can be distributed accordingly.

However, the degree of differences between the delay and delay jitter performances for the lower-priority NRT-type traffic has been shown to be significantly influenced by the burstiness of the NRT-type traffic model. For the ON-OFF NRT traffic model, significant performance differences between the regular CFDAMA and Opt-CFDAMA protocols were observed for the NRT-type. For the SFRP NRT-type traffic model, the burstiness of the NRT-type sources significantly limited the traffic prioritization influence of the LCAS on the Opt-CFDAMA protocols thereby limiting the differences between the Opt-CFDAMA protocols and their corresponding regular variations.

Chapter 6 Conclusions and future research areas

6.0 Conclusions

In the research leading to this thesis, the literature of satellite systems, services and protocols were reviewed. In terms of the applicability of satellite systems to support the demand for communications services in geographically remote areas, we emphasized the role of GEO BSA systems hence we focused our system model, review and analysis on the personal earth station scenario to support residential or SOHO types of end users. Under this situation, the existence of bursty traffic from the end user applications and the requirement for the efficient distribution of the return channel capacity has also been emphasized.

From the developed performance measurement and characterization framework to compare the PR-CFDAMA, W-CFDAMA and the RR-CFDAMA protocol variations, it can concluded that predictive algorithms for estimating traffic trends in the UT's are necessary for efficient capacity utilization in the CFDAMA protocol. However, from the observations of the performance results, it can be inferred that employing predictive algorithms in a BSA system will depend on the characteristics of the UT traffic and the extent of traffic load in the system. At the medium-to-high range of traffic loads, the predictive algorithms are more efficient for a BSA system to support applications that are both delay and delay jitter sensitive since they are more efficient in the capacity distribution under such conditions.

At low-to-medium range, the simple RR-CFDAMA protocol seems to be more efficient for delay and delay jitter sensitive applications, depending on the variability of the UT traffic. If the UT traffic is very bursty (e.g., if the SFRP model is applicable), the RR-CFDAMA protocol will be a better choice of the CFDAMA protocol variations. However, if the UT traffic is known to be not very bursty (e.g. the ON-OFF model), then the predictive algorithms (especially, the proposed EPR-CFDAMA protocol) will be the better choice of the CFDAMA protocols for delay and delay sensitive applications. Moreover, for applications which are delay sensitive but delay jitter insensitive, the predictive protocols will be the better choice of the CFDAMA protocols as can observed from our simulation results.

In Chapter 5, we provided a more detailed strategy for enhancing or optimizing and applying the CFDAMA protocols in the current BSA system based on the IP-DVB-RCS technology. We emphasized the importance of the optimization strategy using application specific information obtainable from the higher-layer protocols and demonstrated it with simulations for an integrated application structure consisting of RT applications employing RTP/RTCP and NRT applications possibly employing TCP.

From the simulation results obtained in chapter five, it can be concluded that the variability in the low-priority component of the integrated traffic structure has significant effects on the QoS differentiation strategy in the CFDAMA protocols. While the delay and delay jitter performances of the Opt-CFDAMA protocols showed marked differences to those of the regular CFDAMA protocol variations for the less bursty ON-OFF NRT traffic model, the same cannot be concluded for the burstier SFRP NRT traffic model.

6.1 Future research areas

The analyses, protocol proposals and methods that have been researched and developed in the research leading to this thesis can all be enhanced to further enable BSA to become fully integrated in current and future communication systems. The area of predictive algorithms for enhancing the performance of the DCA of the predictive CFDAMA protocols need more investigation to further reduce inaccurate capacity allocation and increase system efficiency. While the proposed EPR-CFDAMA protocol has been shown to exhibit measurable and significant performance improvements over the PR-CFDAMA protocol, there are still ample avenues for improvements especially with traffic sources with high variability as exemplified by the SFRP traffic model. A very important aspect of providing cross-layer differentiated services over BSA systems via the CFDAMA protocol is the definition and standardization of the traffic classes that can adequately describe a broad range of application types. In this thesis, we focused on two classes of traffic: the RT and NRT types.

Within each class, there are opportunities for further definitions based on detailed application-specific or intrinsic properties of the UT traffic. Additionally, the distributed QoS provisioning performances of the cross-layer optimized protocols can be enhanced via differentiated traffic prediction, a technique in which each application traffic class in a UT is dynamically predicted as against the aggregated prediction that were employed in our research. While this may increase the protocol complexity, the potential gain of a more accurate traffic prediction warrants such an investigation.

Finally, to enhance the integration of the protocol functions of the CFDAMA protocols in the Internet, further research efforts can be put into the dissemination of slowly-varying cross-layer information to all the UT's and the GES via a systematic signaling technique. The UT's can then have more accurate information about the trends in the other UT's and the GES can use such information for interfacing functions with the terrestrial Internet interface protocols. Further research and performance evaluations can also be carried out to investigate the implementations of the CFDAMA and Opt-CDAMA protocols in the LEO system environment where the propagation delays for the UT's can vary significantly.

Bibliography

[1] S. Kota, M. Goyal, R. Goyal, R. Jain, "Multimedia satellite networks and TCP/IP traffic transport," Borko Furht (Ed.), Internet, multimedia systems and applications (IMSA'99), October 18-21, 1999, Nassau, The Bahamas. IASTED/ACTA Press, 1999, pp. 436-443

[2] D. Bem, T.W. Wickowski and R.J. Zielinski, "Broadband satellite systems", IEEE communications surveys and tutorials, vol. 1, no. 1, First Quarter 2000, pp. 1-15.

[3] T. Le-Ngoc, V. Leung, P.Takats, and P. Garland, "Interactive multimedia satellite access communications," IEEE Communications Magazine, July 2003, pp. 78 – 85.

[4] B.G. Evans, ed., Satellite Communication Systems, London UK: IEE Press ,1999.

[5] M. Allman, ed., "Ongoing TCP research related to satellites," RFC 2760, Feb. 2000.

[6] J. Border et al., "Performance enhancing proxies intended to mitigate link-related degradation," RFC 3155, June 2001.

[7] S. Floyd and V. Jacobian, "Random early detection gateways for congestion avoidance," *IEEE/ACM Trans. Net.* vol. 1 August, 1993 pp. 397-413.

[8] L. Wu, "Dynamic congestion control methods to improve performance of TCP split connections over satellite networks," MASc. Thesis, Univ. of BC, Oct 2002.

[9] J. Stepanek et al., "The use of a proxy onboard the satellite to improve TCP performance," GLOBECOM 2002, Taipei, Taiwan, Nov. 2002.

[10] Digital Video Broadcasting (DVB, interaction channel for satellite distribution systems DVB-RCS EN 301 790 v1.3.1, EBU, CENELEC and ETSI Joint Technical Committee (JTC).

[11] H. Schulzrinne et al., "RTP: A Transport protocol for real-time applications," RFC IETF Audio/Video transport working group.

[12] D.P. Connors, B. Ryu and S. Dao, "Modeling and simulation of broadband satellite networks, Part I: medium access control for QoS provisioning", IEEE Communications Magazine 2000, pp. 128-133.

[13] M. Mobasseri and Victor C.M. Leung, "Bandwidth assignment for VBR traffic in broadband satellite networks", IEEE Communications Magazine 2000.

[14] H. Peyravi, "Medium access protocol performance in satellite communications", *IEEE Commun.* Magazine vol. 37, March 1999, pp. 62-71.

[15] Z. Jiang, and V.C.M Leung," A predictive demand assignment multiple access protocol for broadband satellite networks,"*Int. J. of Satellite Commun.* vol 21 issue 4-5, July-October 2003, pp. 451-467.

[16] P.D. Mitchell, D. Grace, T.C. Tozer, "Comparative performance of the CFDAMA protocol via satellite with various terminal request strategies", *IEEE Communication Magazine*, 2001, pp. 2720-2724.

[17] P.T. Brady, "A Statistical Analysis of ON-OFF Patterns in 16 Conversations", *Bell Syst. Tech. Journal*, pp. 73-91, 1968

[18] A. de Fontenay, M.H. Shugard and D. S. Sibley, *Telecommunications Demand Modeling: an integrated view.* North-Holland: Elsevier, 1990

[19] J.G Gruber, "A comparison of measured and calculated speech temporal relevant to speech activity detection", *IEEE Trans. Commun.*, vol. COM-30, no. 4, pp. 728-738, 1982.

[20] H.M. Hang and J.J. Chen, "Source model for transform video coder and its applications-Part I: fundamental theory," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 7, April 1997, pp. 287-298.

[21] Z. He, Y. Kim, and S. K. Mitra, "A novel linear source model and a unified rate control algorithm for H.263/MPEG-2/MPEG-4," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 3, May 2001, pp. 1777-1780.

[22] D.P. Heyman and T.V. Lakshman, "Statistical analysis and simulation study of video teleconference traffic in ATM networks", *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 2, no. 1, March 1992, pp. 49-59.

[23] R.M. Rodriquez-Dagnino, M.R.K. Khansari and A. Leon-Garcia," Prediction of bit-rate sequences of encoded video signals", *IEEE Journal Selected Areas Commun.*, vol. 9, no. 3, April 1991, pp. 305-313.

[24] P. Jelenkovic, A. Lazar, and N. Semret," The effect of multiple time scale and superexponenentiality in MPEG video streams on queuing behavior", IEEE Journal in selected areas in communications, vol. 15, no. 6, August 1977, pp. 1052-1071.

[25] R. Steinmetz, K. Nahrstedt, "Fundamentals in multimedia computing and communications", Englewood Cliffs, NJ: Prentice Hall, 1995.

[26] D.P. Heyman, "The GRAR source model for VBR video conference", *IEEE/ACM Trans.* on networking, vol. 5, No. 4., August 1997.

[27] O. Lazaro, D. Girma and J. Dunlop, "A study of video source modeling for 3G mobile communications systems", *IEE* 3G *Mobile Communications Technologies Conference*, London UK, pp. 461-465, March 2000.

[28] B. Ryu, "Modeling and simulation of broadband satellite networks- Part II: traffic modeling", *IEEE Commun. Mag.*, vol. 37, no. 3, July 1999, pp. 48-56.

[29] O. Lazaro, D. Girma and J. Dunlop, "Statistical analysis and evaluation of modeling techniques for self-similar video source traffic", *11^{th} IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, vol. 2, pp. 1540-1544, London, September 2000,

[30] C. Huang, M. Devetskiotis, I. Lambadaris, and A.R. Kaye, "Modeling and simulation of self-similar variable bit rate compressed video: a unified approach," Proceedings of SIGCOMM, Cambridge, MA, 1995, pp. 114-125.

[31] M.W. Garret, W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic", Proceedings of SIGCOMM, London, UK, 1994, pp. 269-280.

[32] J. Beram, R. Sherman, M.S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic", *IEEE Trans. on Communications*, vol. 43, no. 2/3/4, Feb./Ma/April 1995, pp. 1566-1579.

[33] F. Yegenoglu, B. Jabari, and Y. Zhang, "Motion-classified autoregressive modeling of variable bit rate video," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 3, 1993, pp. 42-53.

[34] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *Proc. ACM SIGCOMM* 1994, London, UK, pp. 257-268.

[35] A. Karasaridis, D. Hatzinakos, "Network heavy traffic modeling using α-Stable self-similar processes", *IEEE Trans.* on Communications, vol. 49, no. 7, July 2001, pp. 1203-1214.

[36] M.E. Crovella, A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes", *IEEE/ACM Trans.* on networking, vol. 5, no. 6, Dec. 1997, pp. 835 -845.

[37] B. K. Ryu and S.B. Lowen, "Point process approaches to the modeling and analyses of self-similar traffic- Part I: model construction", *IEEE/ACM Trans.* on networking, 1996, pp. 1468-1475.

[38] W. Willinger, M.S. Taqqu, W.E. Leland, and D.V. Wilson, "Self-similarity in high-speed packet traffic: analyses and modeling of Ethernet traffic measurements", *Statistical Science*, vol. 10, no. 1, 1995, pp. 67-85.

[39] S. Blake et al., "An architecture for differentiated services", RFC 2475, 1998.

[40] T. Le-Ngoc and I.M. Jahangir, "Performance analysis of CFDAMA-PB protocol for packet satellite communications", *IEEE Trans.* on *Commun.*, vol. 46, no. 9, Sept. 1998, pp. 1206-1214.

[41] I.M. Jacobs, R. Binder, and E.V. Hoverstein, "General purpose packet satellite networks," *IEEE Proc.*, vol. 66, No. 11, 1978, pp. 1448-1467.

[42] M. Hasegawa, G. Wu, and M, Mizuno, "Application of non-linear prediction methods to the Internet traffic", *Proceedings of the IEEE International symposium on Circuits and Systems* vol. 3, Sydney, Australia 2001, pp. 169-172.

[43] B. Ryu, Fractal network traffic: from understanding to implications, PhD thesis, Columbia University, 1996

[44] T. Le-Ngoc and S.V. Krishnamurthy, "Performance of combined free/demand assignment multiple access scheme in satellite communications," *International Journal of Satellite Commun.*, vol. 14, pp.11-21, 1996.

[45] J. Mohammed and T. Le-Ngoc, "Performance analysis of combined free/demand assignment multiple access (CFDAMA) protocols for packet Satellite communications," Proc. ICC'94, New Orleans, LA, pp. 869-873, May 1994.

[46] www-tkn.ee.tu-berlin.de/research/trace/trace.html

[47] M. Mobasseri, *Performance evaluation of* MAC *protocols and buffer management strategies in* ATM-*based broadband satellite networks*, Master's thesis, University of British Columbia, June 2000.

[48] Z. Jiang, Y. Li and V.C.M. Leung, "A predictive demand assignment multiple access protocol for broadband satellite networks supporting Internet applications," IEEE ICC, May 2002 pp. 2973-77.