Performance Analysis of Broadband Multimedia Wireless Communication Networks

Thimma V.J. Ganesh Babu

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy at
Concordia University
Montreal. Quebec. Canada

October 2001

0-612-63989-4

ABSTRACT

Performance Analysis of Broadband Multimedia Wireless Communication Networks

Thimma V.J. Ganesh Babu, Ph.D.,
Concordia University, 2001.

The designing of broadband multimedia wireless network systems should aim at achieving maximum utilisation of wireless resources through statistical multiplexing, while. at the same time satisfying the Quality of Service(QoS) requirements of multimedia traffic. In this research, we consider a priority based scheduling strategy. suitable to the terrestrial/satellite wireless environment. The multimedia traffic is categorised into real-time (voice and video connections) and non-real-time(data connections) depending on whether it is delay sensitive or loss sensitive. The fixed size packets generated by each of the aggregated voice, video and data sources from all user terminals in an uplink beam, are modeled as a 2-state Markov modulated Poisson Process (MMPP). Using the counting process of real-time traffic, the real-time packet loss probability has been evaluated at the uplink. Based on the equation governing the non-real-time packet queueing process at the epochs of the beginning of each frame and by using an embedded Markov chain analysis, the elements of the transition probability matrix are derived. Using the matrix-geometric technique, the occupancy distribution non-real-time packet queue is evaluated. The illustrative results for different cases of traffic mix are presented. Further, we outline the analytical derivation for obtaining covariance function of number of real-time and non-real-time arrivals to a particular downstream link through the switch. We match the covariance function values at different lags with the covariance function of 2-state MMPP at corresponding lags in order to obtain the parameters of approximating 2-state MMPPs. Based on this, and using the single queue model of the uplink, we describe the procedure for evaluating the performance at the downlink. A simulation model has also been developed in order to assess the effects of the various approximations required for the analytical model.

Acknowledgement

I express my sincere and deepest gratitude to both of my thesis supervisors Prof. Tho Le-Ngoc and Prof. Jeremiah F. Hayes. Although the two eyes of our face. are spatially at different co-ordinates, but still they help to focus on the same single scenario. In the same manner, both professors have been helping me to focus on the single objective of performing an exciting and excellent research work. The discussions I had with both professors not only helped me to have the intellectual maturity but also the philosophical maturity.

I wish to sincerely thank Prof. Mehmet Ali for providing feedback and suggestions while doing this research work. The discussions I had with Prof. Mehmet Ali during the course work provided me with the acquisitions of all the relevant fundamentals for doing my research work. I also would like to thank Prof. J. W. Atwood for having discussions on practical aspects of research. Also, I would like to express my gratitude to Prof. Reza Soleymani, for having many discussions during my research work.

I express my sincere gratitude to the CITR for providing me with financial support for doing this research work. In particular the help extended by the industrial counterparts from SPAR Aerospace Ltd., by having discussions and providing many practical details, is highly commendable. In this regard, I would like to express my deepest gratitude to Dr. Augustin Iuoras and Mr. Martin Cote for having discussions.

I would like to express my deepest gratitude to Prof. Marcel F. Neuts, for helping me by providing suggestions and appropriate references to the on-going research

work which were extremely helpful for my research.

I am very grateful to the Graduate Secretary Madame Diane Moffat for the timely arrangements of all the official procedures needed to complete my thesis work. I would like to express my indebtedness to my family who stood by me and provided their support while doing this research. Finally I would like to thank all my friends who made my stay and my studies quite exciting and enjoyable.

Contents

List of Figures

List of Tables

List of Abbreviations

ATM	Asynchronous Transfer Mode
BMAP	Batch Markovian Arrival Process
CDMA	Code Division Multiple Access
CFDAMA	Combined Free and Demand Assignment Multiple Access
CLR	Cell Loss Ratio
CTD	Cell Transfer Delay
DTH	Direct-To-Home
DAMA	Demand Assignment Multiple Access
EIRP	Equivalent Isotropically Radiated Power
FDD	Frequency Division Duplexing
FDMA	Frequency Division Multiple Access
FIFO	First In First Out
G/T	Gain to noise temperature ratio
IDC	Index of Dispersion for Counts
LEO	Low Earth Orbit
GEO	Geostationary Earth Orbit
MAC	Medium Access Control
MEO	Medium Earth Orbit
MMPP	Markov Modulated Poisson Process
NRT	Non-Real Time
OBS	On-Board Switching
PODA	Priority Oriented Demand Assignment
QBD	Quasi-Birth Death
QOS	Quality Of Service
RT	Real Time
SMSP	Semi-Markov Service Process
SMVP	Semi-Markov Vacation Process
SMG	Statistical Multiplexing Gain
TDD	Time Division Duplexing
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
VSAT	Very Small Aperture Terminal

List of Symbols

N_{voi}, N_{vid}, N_{data}	Number of on-going voice, video and data calls (Page # 27)
n_0	Number of transient samples (Page # 36)
X_i	Sample value of the sequence from simulation (Page # 36)
$\hat{\rho}(k)$	Sample autocorrelation function (Page # 36)
$\hat{\gamma}(k)$	Sample covariance function (Page # 36)
$\pm\mathbf{D}$	The limit within which long term $\hat{\rho}(k)$ lies (Page # 36)
L	The number of samples after which $\hat{\rho}(k)$ lies within $\pm\mathbf{D}$ (Page # 36)
N	$N = J * L$, where J is any positive integer (Page # 36)
N'	$N' + n_0$ is number of samples generated from single run (Page # 36)
M	equals $\frac{N'}{N}$ different number of batches from simulation (Page # 36)
X_{mn}	n-th sequence value of batch m (Page # 36)
v_{im}	number of X_{mn} in batch m whose value is i (Page # 36)
$\hat{p}_{im}$	equals $\frac{v_{im}}{N}$ (Page # 36)
p_i	equals $\hat{E}(\hat{p}_{im})$ (Page # 36)
$\hat{p}_i$	The estimator of probability of queue occupancy being equal to i (Page # 37)
$s^2(\hat{p}_{im})$	The variance of the random variable $\hat{p}_{im}$ (Page # 37)
$t_n(x)$	The $100x$-th percentile of t-distribution with n-degrees of freedom (Page # 37)
$q(\tau_n)$	The queue size at time $t = \tau_n$ (Page # 37)
$A_k(\tau_n, \tau_{n-1})$	The random variable representing number of arrivals during the period (τ_n, τ_{n-1}) (Page # 37)
C_n	random variable representing the number of units from the buffer that can be served during the interval (τ_n, τ_{n-1}) (Page # 37)
$(N(t), J(t))$	the joint state of queue and phase of input process (Page # 38)
Q	Infinitesimal generator matrix of the process $(N(t), J(t))$ (Page # 38)
D_k	the non-zero submatrices infinitesimal generator matrix of Q (Page # 38)
$\tilde{H}(x)$	Distribution function of service times (Page # 39)
$H(s)$	The Laplace transform of $\tilde{H}(x)$ (Page # 39)
ρ	The traffic intensity (Page # 39)
$(X_n, J_n, (\tau_n, \tau_{n-1}))$	the state of queue, BMAP phase at time τ_n and period (Page # 39)
$\tilde{P}(x)$	The transition probability matrix of the process $(X_n, J_n, (\tau_n, \tau_{n-1}))$ (Page # 39)

$\tilde{A}_n(x)$ and $\tilde{B}_n(x)$	See Page # 39
$\tilde{G}_{jl}^{[1]}(k;x)$	See Page # 39
$\tilde{G}_{jl}(k;x)$	See Page # 40
$G(z,s)$	See Page # 40
$A_v(s)$	See Page # 40
G	See Page # 40
$\mathbf{g}$	See Page # 40
$\mathbf{P}$	See Page # 40
$\mathbf{x}_i$	See Page # 40
$q(n), M$	See Page # 42
$p_{m,n}$	See Page # 42
p_{ij}	See Page # 42
$p_{kl,ij}$	See Page # 42
$Q_j(z), Q(z)$	See Page # 43
$B_j(z), B(z)$	See Page # 43
$g_i(z), h_i(z), \lambda_j(z)$	See Page # 43
$q'(t), C$	See Page # 45
$N_1, N_2, A_1, A_2, a, b, c, d$	See Page # 45
$F_{ij}(y), \mathbf{F}$	See Page # 46
$Q_{vo}, Q_{vi}, Q_r, \sigma_{1,vo}, \sigma_{2,vo}, \sigma_{1,vi}, \sigma_{2,vi}$	See Page # 53
$\Lambda_{vo}, \Lambda_{vi}, \Lambda_r, \lambda_{1,vo}, \lambda_{2,vo}, \lambda_{1,vi}, \lambda_{2,vi}$	See Page # 54
$\pi_r, \underline{e}$	See Page # 54
$\tau, N_r(n\tau), J_r(n\tau), J_d(n\tau)$	See Page # 55
$P_{r_{ij}}(k, n\tau), P_r(k), f_r(k), CLP_{RT}$	See Page # 55
$N_d(n\tau), q(n\tau), P_{d_{ijl,uvw}}, P_d$	See Page # 56
$G_i, P_r(m), P_d(i)$	See Page # 57
D, U_i	See Page # 58
E_i, A_0, A_1	See Page # 59
$A_i, B_i, \Pi_{uvw}, \Pi_d, \underline{1}, l, m, n$	See Page # 60
$\underline{\beta^*}, \Delta(\underline{\beta^*}), A, \underline{\kappa}, \underline{\phi}$	See Page # 61
C_u	See Page # 62
$P(N_r'(t) = k, J_r(t) = j/J_r(0) = i)$	See Page # 78
$P(N_d'(t) = k, J_d(t) = j/J_d(0) = i)$	See Page # 78
$N_{a,RT}(i\tau), N_{d,RT}(i\tau), N$	See Page # 79
$m1, i1, m2, i2, Cov_{ind}(N_{a,RT}(n\tau), N_{a,RT}((n+j)\tau))$	See Page # 82
$N_{a,NRT}(i\tau), N_{d,NRT}(n\tau)$	See Page # 82
$Cov_{ind}(N_{a,NRT}(n\tau), N_{a,NRT}((n+j)\tau))$	See Page # 90
$N(n\tau), \sigma_1, \sigma_2, \lambda_1, \lambda_2$	See Page # 91
$r, Cov_{tot}(N_{NRT}(n\tau), N_{NRT}((n+1)\tau))$	See Page # 91

Chapter 1

Introduction

1.1 Multimedia Services on Wireless Networks

The concept of telecommunications has evolved from providing simple telephone communication services to versatile multimedia services. Because of the transition from analog communications to digital communications. variable bandwidth requirements of multimedia services can be provided by a single ubiquitous communication network for mobile. fixed or home users. The ITU (International Telecommunications Union). has developed an unified architecture called IMT-2000 (International Mobile Telecommunications-2000). for supporting ubiquitous communication from anywhere in the world. Wireless communication becomes an essential part of this architecture. as providing wirelines to users in extremely remote places becomes economically infeasible. For this purpose. IMT-2000. includes both satellite based and terrestrial wireless based communication in its architecture [12]. to provide global coverage. ITU-T (ITU-Telecommunications Sector). defines the networks aspects. while ITU-R (ITU-Radiocommunications Sector) defines the radio aspects of this architecture. Recently. ITU has been developing standards for the Global Information Infrastructure (GII) [13]. for providing global interconnectivity and interoperability.

In United States. the Satellite Communication Division (SCD) of Telecommunications Industry Association (TIA). is working on providing standards for satellite communications, with a focus on interoperability and spectrum issues. The SCD organises its operation, through the TR-34 Committee, Satellite Equipment and Systems, and its related subcommittees, namely, TR-34.1 Communications and Interoperability, and TR-34.2 Spectrum and Orbit Utilisation. Engineering Subcommittee TR-34.1, in conjunction with the Communications and Interoperability Section (CIS) works on various issues and problems of interoperability between and among terrestrial and satellite networks. This subcommittee is mainly focusing on GSM, ATM and Internet and their seamless operation over satellite networks. It has defined a set of satellite based ATM network architectures for future physical layer specifications. The architectures defined by TR-34.1 can be broadly grouped into two categories: Satellite ATM (SATATM) architecture for transparent (bent-pipe) satellites and Satellite ATM (SATATM) architecture with on-board switches [13].

The Satellite networks themselves can be Geostationary Earth Orbit (GEO), Medium Earth Orbit (MEO) or Low Earth Orbit (LEO) based constellations or combinations thereof. Any satellite architecture has two basic components, namely, earth segment which includes user terminals/earth stations and network control center, and space segment which has the complementary hardware and software components on the satellite [21] [22]. There are three types of commonly deployed Ka-band satellite network architectures namely, bent pipe, multiple spot beams with inter-beam connectivity provided by terrestrial networks and multiple spot beams with inter-beam connectivity provided by on-board switch [5]. Earlier satellite systems provided Narrowband Integrated Services Digital Network (N-ISDN) capabilities to Very Small Aperture Terminals (VSAT) [14] through a "bent pipe" transponder with the switching of traffic performed by the central hub on the ground. But, the new generation satellite systems use mutiple high gain spot beams with on-board switching and processing capabilities to provide broadband services. The multibeam configuration with on-board switching has the following advantages over the central hub based architecture.

- Because of the use of the Ka frequency band (20/30 GHz), and the high EIRP and G/T of satellite antenna due to significantly reduced coverage of the high gain spot beam, both the antenna size of earth/user terminal and the power of the transmitted signal can be made smaller. Thus, this multibeam configuration allows the use of small, inexpensive user terminals. This enables providing Direct-to-Home (DTH) multimedia services at a reasonable cost.

- Easy configuration of private virtual networks, since better traffic isolation is possible.

- The link performance is improved due to on-board demodulation and remodulation, since uplink and downlink transmissions can be optimised individually.

Since satellite systems are power, bandwidth and weight limited, the design of space segment which involves the design of RF transponder, on-board control components, on-board memory size, etc., should be carefully performed. This care in design is reflected in the ground segment as well. Tradeoffs must be made with respect to the complexity, power and weight requirements for providing on-board buffering, switching and processing features of the satellite network. A large number of commercial systems (Teledesic (USA), SPACEWAY (USA), ASTROLINK (USA), CYBERSTAR (USA), CELESTRI (USA), N-STAR (Japan), WEST (EUROPE), EUROSKYWAY (EUROPE), SKYBRIDGE (EUROPE)) [15] [16] have been proposed from all over the world at the Ka frequency band, targeting multimedia services at high data rates mainly to fixed and possibly to mobile user terminals.

The on-board switching based multibeam satellite network architecture, can be either circuit switching based architecture or packet switching based architecture or a combination of both (hybrid architecture) [4]. Intermediate between the two there is also burst switching based architectures which is basically circuit switching at the burst level [76][77][89][90]. The issues involved in burst switching are similar to circuit switching (in terms of implementation complexity). Although the circuit switching based architecture is simpler to implement, compared to packet switching based architectures, it results in poor utilisation of satellite capacity. This is because most traffic types are bursty and have associated peak bit rate and average bit rate. The circuit switching based architecture allocates a circuit, i.e., a channel at the uplink (from user terminals to the on-board switch) and the downlink (from the on-board switch to the user terminals) with the capacity equal to peak bit rate to such traffic types and during the times when the source of such traffic types, is not generating traffic at the peak bit rate, the unused residual capacity cannot be reallocated to any other traffic source and thus goes to waste. On the other hand, with packet switching based architectures, due to statistical multiplexing, allocated capacity is relative to the average throughput requirement. The traffic information is packetised and queued for transmission until the capacity is available. Thus, this results in higher utilisation of capacity, although there are additional issues of designing suitable protocols, traffic, congestion and admission control, to be taken care of. In particular, because of the statistical multiplexing, the packets experience stochastically varying delay due to buffering and loss when the buffer is full. Moreover, some traffic types cannot tolerate more than certain amount of delay and loss: the tolerance of such traffic types is specified in terms of required quality of service. Of course, with peak rate allocation of channel capacity in the case of circuit switching based architecture, there is no loss and delay of information apart from propagation delay. The main factor involved in choosing packet switching based architecture against circuit switching based architecture, is the improved utilisation of the satellite capacity due to the statistical multiplexing. The statistical multiplexing gain (SMG) in the case of packet switching based architectures, can be defined as the ratio of sum of product of number of each traffic type and its corresponding peak rate to the satellite capacity for a given allowable loss probability and delay performance.

The important issues here are the QoS to be provided to the traffic and maximising the utilisation of the satellite network resources whether the satellite network is stand-alone or acting as a subnetwork for interconnecting terrestrial gateways/networks. In order to design schemes to address these issues, it becomes essential to understand different components of the satellite network where such schemes will be implemented. When the multiple user terminals need to send traffic to the on-board switch on the uplink beam, either they can send traffic by randomly accessing the uplink or they can send requests and then send their traffic data on the uplink when the MAC scheduler allocates uplink slots, such as PODA, CFDAMA [19][20], etc. On the uplink beam, the Medium Access Control (MAC) used can be, based on FDMA, multiple frequency time division multiple access (MF-TDMA) or CDMA and on the downlink, the access is usually based on simple time division mutliplexing (TDM). Once the traffic arrives at the on-board switch, it has to be routed to the appropriate downlink. At the on-board packet switch, contention can occur when two or more packets at the input need to be routed to the same output. Thus, there are two types of packet switching architectures available in practice, one based on contention resolution and the other based on contention-free switching. Once the packets are routed to the appropriate downlink, the packets are scheduled for transmission on the downlink TDM by the downlink scheduler. Therefore, ensuring a strict QoS control, over the QoS provided to the traffic originating at the user terminals, can only be done by the appropriate scheduling function at the MAC scheduler and Downlink scheduler, when the switch is non-blocking (we refer to non-blocking switch as ideal switch).

This scheduling function and on-board switching function can be seen as MAC layer functions which are parts of layer 2, Data Link Control (DLC) of the seven layer protocol stack proposed by ISO/OSI standard. The higher layer standard protocols such as IP or ATM can be using this MAC layer functionality with appropriate SNDCF (subnetwork dependent convergence functions). Therefore, standard applications using internet, such as ftp, telnet, VOIP (voice over IP), etc., can be supported. For the fibre-optic based wireline communication, the ATM concept has been proved to be quite successful. In providing global information infrastructure, satellite and/or terrestrial wireless networks would be required to support ATM connection services, although the limitation of long propagation delay in the case of satellite network and the limitation of low bit rate (as compared to Gigabit support of single fibre-optic channel) in the case of terrestrial wireless networks, exists. ATM networks meet the flexible bandwidth and quality of service (QoS) requirements of multimedia services, by statistically multiplexing fixed-size packets of different traffic types. Recently, there has been a great interest in extending these capabilities to satellite and terrestrial wireless networks. There are many projects being carried out for implementing ATM over Satellite as outlined in [7]. In particular, there is great interest in having ATM on the return link and DVB-S (Digital Video Broadcasting over Satellite) on the forward link of the satellite system.

The ATM Forum has been developing a set of functional specifications for terrestrial Wireless ATM (WATM), including Mobile ATM (MATM) for mobility support within an ATM network and Radio Access Layer (RAL) for ATM-based terrestrial wireless access [67]. This WATM group has also done work on ATM over satellite [7]. As part of WATM radio specifications, ATM Forum has also issued specifications for using the physical satellite channel. There are many WATM projects being carried out worldwide, such as Olivetti Research Laboratory's (ORL) Radio ATM (U.K.) [62], NTT AWA (Japan) [63], The Magic WAND (Wireless ATM Network Demonstrator in Europe) [23][24][65], NEC C&C Research Laboratories (U.S.) [26].

Thus, the important issues here, once again, are the QoS to be provided to the traffic and maximising the utilisation of terrestrial wireless network resources. When the multiple user terminals need to send traffic to the base station, either they can send traffic data by randomly accessing the uplink like in PRMA and its variants, or they can send requests and then send their traffic data on the uplink, when the base station scheduler allocates uplink slots, such as in Dynamic Time Division Multiple

Access (DTDMA). Once the traffic is received by the base station, these are routed by the switch to the appropriate downstream link. The scheduler at the downstream links will again schedule the transmission of these traffic. Therefore, ensuring a strict control over the QoS provided to the traffic originating at the user terminals, can only be done by the appropriate scheduling function at the base station and at downstream links.

1.2 Focus of Research

The multimedia traffic types can in general be classified as real-time traffic (voice, video, etc.) and non-real-time (bursty data) traffic. The satellite OBS environment is quite different from the terrestrial packet switching system in two aspects: firstly, from the time a packet arrives and the transmission request is sent until its transmission at the uplink, the packet has to wait for at least one round-trip propagation delay. Secondly, since the satellite is weight limited, the buffer available for queueing packets at the downlink is limited. The first factor will result in larger queueing delay at the earth stations in addition to the large propagation delay as compared to the delay involved in an equivalent terrestrial multiplexer. The second factor could lead to higher packet loss probabilities due to limited buffer size, as compared to the case of terrestrial switching system, where the buffer size is not critical.

Our main focus is on evaluating the QoS in terms of CLR of real-time traffic and CLR of non-real-time traffic at the uplink and characterising the arrival process of both of these traffic types at the downstream links so that the QoS in terms of CLR of real-time traffic and non-real-time traffic can also be evaluated at the downstream link. The scheduling function that we consider is priority based with priority given to real-time traffic over non-real-time traffic.

In [6], only jitter-tolerant traffic was considered, the mean end-to-end delay was evaluated by individually evaluating the mean delay of messages at uplink and at on-board switch (both cases of input-queueing and output queueing were considered). The drawback associated with this analysis is that it does not consider real-time traffic and accordingly no sophisticated scheduling is considered for providing QoS differentiation between real-time traffic and jitter-tolerant traffic. In [8], modeling voice sources as on-off sources, simulation results of CLR for voice sources and CLR for ABR traffic with fixed buffer size for a link capacity of 16 Mbps were presented. In [9], from each user terminal, it is assumed that there are isochronous sources (stream type) admitted, which are handled on a circuit switching basis and during the times when the isochronous sources become inactive. the bursty data traffic can be sent on those empty channels. The user terminal loss performance of data sources at the uplink is presented for different data buffer sizes based on stochastic fluid flow model and simulation. To analyse the loss performance of data traffic at on-board packet switch, the stochastic fluid flow model along with a renewal approximation was used and analytical results of loss performance for different buffer sizes were compared with simulation assuming uniform distribution for destination of data packets. In [10], the isochronous (CBR) trafffic and bursty data traffic have their own reserved slots and a common resource pool slots in a TDMA frame. Both isochronous (CBR) traffic and bursty data traffic are allowed to use common resource pool if available and this scheme is called Double Movable Boundary scheme. CBR calls are assumed to arrive according to Poisson distribution and their holding time is defined as the number of frames which is geometrically distributed and data packets are assumed to arrive according to Poisson distribution in a frame period. An embedded Markov chain analysis at the end of frame periods is developed to analyse the data queue. Simulation results with MMPP data sources were also presented. As compared to the analysis presented in [9], we can see that CBR calls are assumed to be active during the holding time.

Next, we discuss the models of resource allocation proposed for wireless terrestrial networks. In [25], for different average burst sizes of ABR, connections based on the contention resolution mechanism (using time division multiple access/time division duplexing (TDMA/TDD)), emulation results (using a software package) of the mean ABR cell delay vs. throughput. assuming a fixed number of ABR connections in progress, were presented. Also the achievable throughput for ABR/CBR mix was investigated with circuit-oriented service given to CBR Virtual Connections (VC's). However, no VBR traffic was considered. In [64], a static priority assignment for real-time (rt) VBR VC's was made after considering an optimization problem to minimize the overall CLR, with CLR constraints on the individual VC's. This optimization problem assumes a M/D/1/FCFS/Nonpre queue. The CLR for a particular VC is assumed to be the probability of exceeding the maximum tolerable cell delay of that VC. Also the optimization problem of a deadline-oriented dynamic priority discipline called relative urgency policy. as applied to a set of VBR VC's assuming a Poisson arrival process. were presented. The main problem associated with this technique is the computational complexity, which increases exponentially with the number of VCs. Further. the Poisson cell arrival process assumption for VCs is not truly representative of real traffic. Simulation results assuming autoregressive arrival process of VC cells were also presented in order to compare static priority based allocation with dynamic priority allocation based on relative urgency policy. In [65], DSA++ MAC scheme was assumed. Based on static priorities with priority order of CBR > VBR > ABR and relative urgency based scheduling within CBR and VBR connections, simulation results of mean delay and CDF of delays for different mixes of CBR voice, VBR video and ABR data traffic, were presented. In [66] a static priority based allocation policy was considered to support CBR, VBR and ABR sources. The slot allocation for CBR and VBR sources was done based on polling. The polling parameters are chosen such that the delay constraints of all CBR and VBR sources can be satisfied based on a recursive scheme. The slot allocation for ABR sources was done based on group randomly addressed polling. The problem associated with this method is that the deterministic bound on QoS for CBR and VBR sources (assuming a deterministic traffic characterization with a leaky bucket mechanism) is too conservative compared to the probabilistic QoS obtained through simulation. For this simulation, the VBR sources were assumed to be On-Off sources. Also, to perform connection admission control (CAC) based on this scheme, extensive computation is required to examine whether the new admitted connection would violate the delay constraints of existing sources.

Thus, these previous studies were carried out either by simulation or by analyses based on simple (uncorrelated) traffic assumptions. Our work is motivated by the need for computing achievable QoS under realistic bursty traffic assumptions. Thus, our work differs from them in two main aspects. The traffic is modeled as MMPP's for rt-VBR and non-real-time (nrt) VBR connections. We can obtain the actual performance metric of interest by direct analytic computations, based on the matrix-geometric technique. Our analysis can be used to evaluate the performance at any downstream link, since we model the arrival process to any downstream based on the departure process from the upstream links.

Thus, our objective is to evaluate the CLR of real-time traffic and CLR of non-real-time traffic at the uplink assuming MMPP traffic models for both types of traffic and to evaluate the CLR of real-time traffic and CLR of non-real-time traffic at the downstream link by approximating the arrival process of real-time traffic and non-real-time traffic arriving to the downstream link through the switch, by MMPPs.

We enlist the contributions made from our research work as follows:

- We developed an embedded Markov chain analysis for evaluating the queue occupancy of non-real-time data queue using matrix-geometric technique, with realistic traffic models and with priority based scheduling.

- We observe that the evaluation using matrix geometric technique becomes much more simplified compared to the general case due to the case of $A_{i+1} = B_i$ for $i \geq 1$.

- We developed a matrix product based methodology to evaluate the covariance function of real-time traffic arrival process and non-real-time traffic arrival process to a downstream link assuming symmetric traffic loading conditions at the uplinks and uniformly distributed destination at the switch.
- By using the covariance function values at different lags and mean arrival rate. we approximate the actual arrival process of real-time traffic and non-real-time traffic by 2-state MMPPs and using the embedded Markov chain analysis developed for the uplink. we evaluate the CLR of real-time traffic and survivor function of non-real-time queue at the downstream link.

The outline of the thesis is as follows. In Chapter 2. we discuss the broadband wireless system configurations. giving particular emphasis to Satellite ATM and typical configurations of terrestrial wireless ATM systems. In Chapter 3. after describing the scheduler structure with priority based scheduling, the problem statement is given. Following this. the problem details of uplink and downstream link with non-blocking type of switch used between uplinks and downstream links are discussed. In Chapter 4. performance evaluation techniques based on matrix-geometric method. generating function method and fluid flow approximation method. available in the literature are discussed. Using the matrix geometric method is justified in Chapter 4. In Chapter 5. we discuss the mathematical model of priority based scheduling for the uplink multiplexing system and present the results for different mixes of traffic at different loads. In Chapter 6. we discuss the characterisation of traffic arriving to a downstream link and describe how this characterisation is used to analyse the performance at the downstream link. The main difference between the analyses of uplink and downlink, is that at the uplink. we use approximation by MMPPs for the superposition of On-Off/mini-On-Off source models. while at the downlink. due to multiplexing from different uplinks. we have to model the arrival processes as MMPPs. by appropriately characterising the net multiplexed traffic arriving at the downlink through the switch. In Chapter 7, we present our conclusions and future work. In Appendices A and B, we present the details of the simulation models.

Chapter 2

System Configurations

A typical on-board packet switching based broadband satellite communication system is as shown in Figure 2.1. In general. the traffic from the satellite user terminals can be categorised as real-time and non-real-time traffic. Real-time traffic means the type of traffic whose delay performance requirement is most critical while moderate packet loss can be tolerated. Similarly. non-real-time traffic means the type of traffic that is sensitive to packet loss. while moderate delay can be tolerated. The real-time and non-real-time connections originate at the user terminals. From the users' point of view. these connections should get better than some minimum QoS. From their point of view QoS is very important. On the other hand. the network controller/operator is striving to maximise the utilisation of the system by trying to admit as many connections as possible at the same time providing the guaranteed QoS to the admitted active connections. When more connections can be admitted into the system more revenue can be made. This can only be achieved by the efficient scheduling function. which should allow for statistical multiplexing of the packets. and provide the required QoS to the admitted active connections. The typical QoS parameters can be the probability of the delay exceeding some threshold value is less than a particular given value. and the probability of loss does not exceed a particular given value with the fixed buffer size. For this purpose. we consider a priority based scheduling with priority given to real-time traffic over non-real-time traffic in providing the network

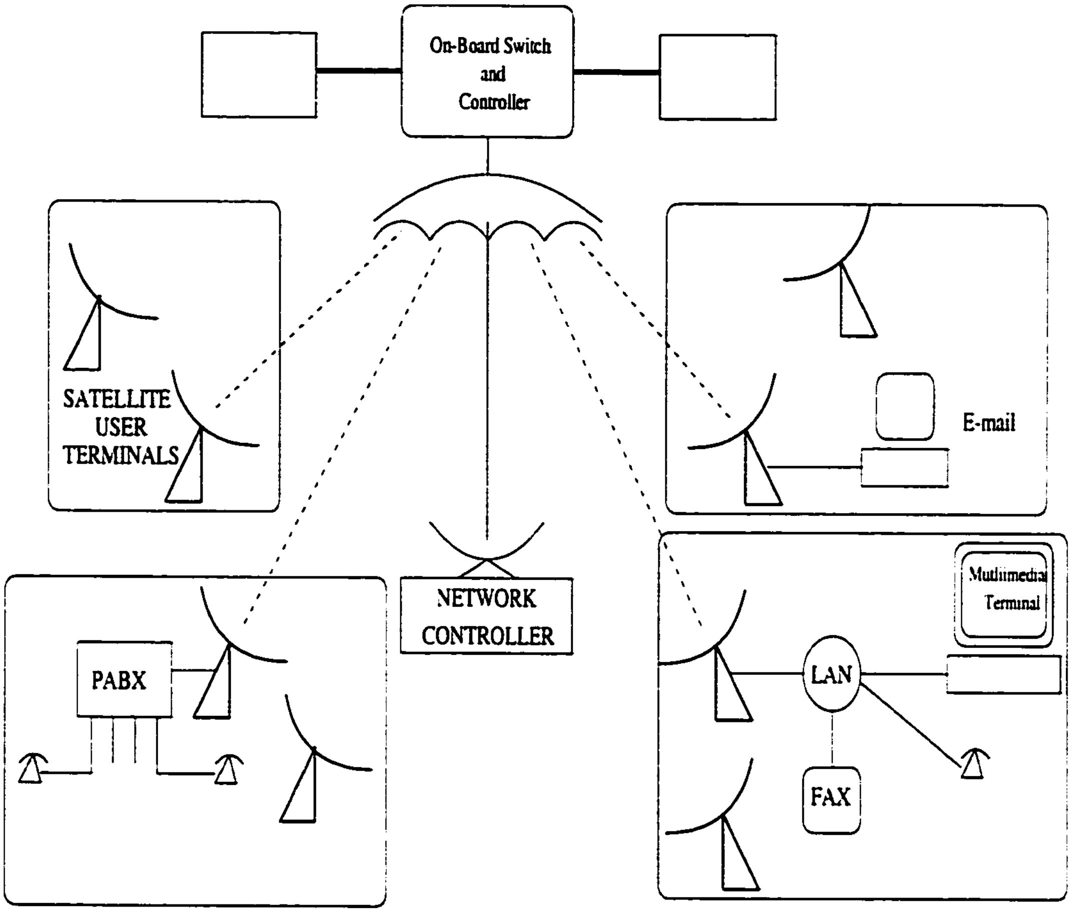

Figure 2.1: The Satellite Communication System with an On-Board Packet Switch.

resources. This is because the real-time connections have very stringent requirements on delay.

2.1 Satellite ATM

Supporting ATM based transfer across such a system can be performed in two ways. The protocol suite across the satellite network can be either based entirely on the ATM structure with appropriate modification to be performed at the ATM layer or it can be based on protocol encapsulation (or interworking) with the appropriate Data link control and network control for transferring ATM cells. The typical Protocol stacks for both control plane and user plane are shown in Figure 2.2 and in Figure 2.3 for these respective approaches. The salient difference between these two protocol stacks is the data link functionality of error control and connection management at the data link level, which are parts of functionalities of LLC (Sat) as shown in Figure 2.3 but these functionalities would not be implemented in Figure 2.2.

There are two types of terminals connected to the satellite network, namely User Terminals and Gateway Terminals (or Gateway Stations) apart from the Network Control Station. In general, the Gateway Stations provide connectivity to terrestrial ATM networks to the User Terminals. At the user plane, the ATM virtual connections are admitted into the satellite network, only if there is enough resource available to satisfy the required quality of service, in terms of loss and delay. The control plane for this purpose will require extended functionality in addition to control functions supported for terrestrial networks. In particular, when satellite User Terminals are mobile, the additional control functions associated with mobility need to be supported.

2.2 Wireless ATM

WATM implementation can be performed in two ways. It can be based entirely on ATM structure with the additional error control and resource control through MAC scheduling in order to provide QoS guarantees at the ATM connection end points as a single ATM network. Alternatively, it can be based on inter-working, which processes AAL Protocol Data Units (PDU's) in an optimized manner suitable for the wireless environment [67]. The configuration of the WATM protocol suite, which transports ATM cells transparently across the wireless medium through the wireless MAC and to the ATM switch, is shown in Figure 2.4. A typical architecture of WATM systems is shown in Figure 2.5. The user-terminals in a cluster share the radio channel to transmit/receive ATM cells to/from the base-station. A base-station

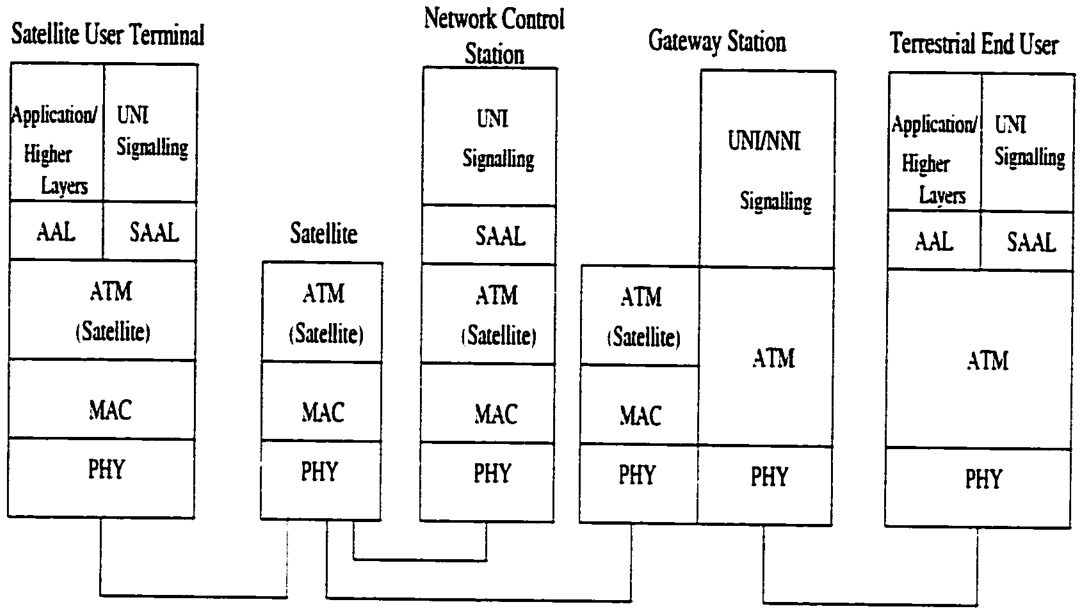

Figure 2.2: Native ATM based protocol suite of Satellite Network with On-board Switching.

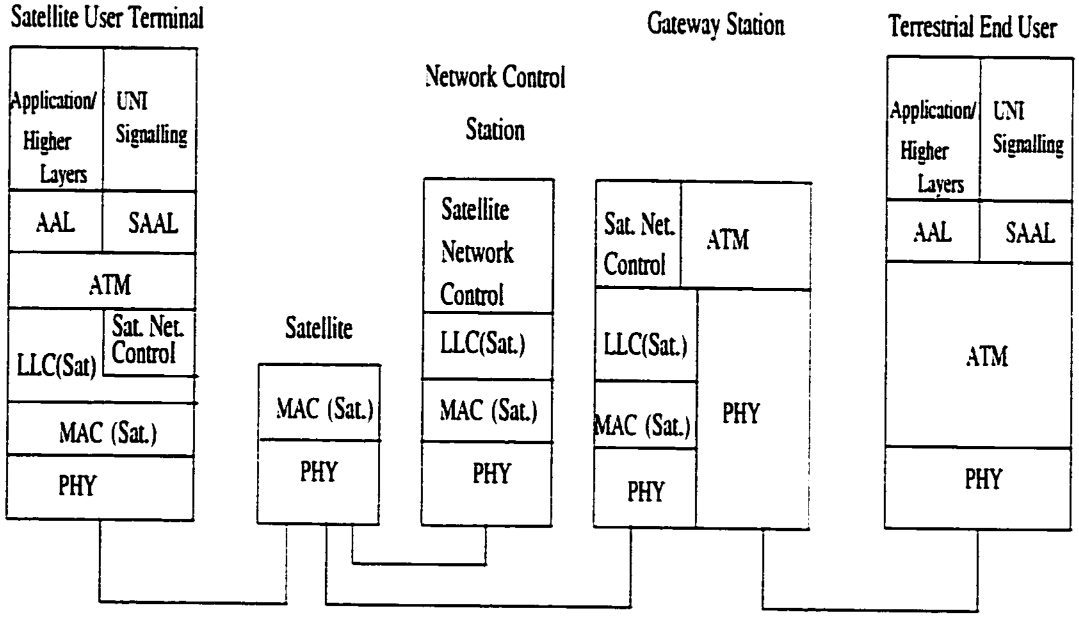

Figure 2.3: ATM based protocol suite with Protocol Encapsulation provided by a Satellite Network with On-board Switching.

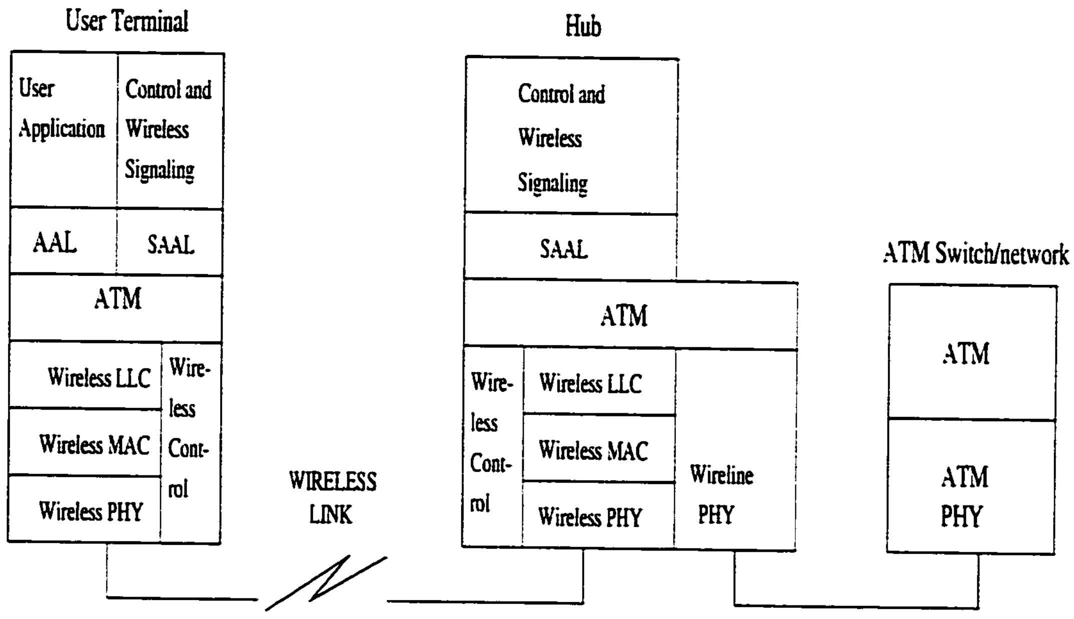

Figure 2.4: ATM based protocol suite of user terminal. hub and ATM switch.

can be connected directly to an input port of an ATM switch (which is assumed to be non-blocking and output queued) or via an ATM multiplexer. The traffic received from the remote terminals by the base station in a cluster is passed to one input port of the ATM switch (either directly or after being multiplexed with traffic from other base stations) and traffic from an output port of the ATM switch is passed to either the corresponding base station. which in turn distributes the traffic to the remote terminals, or to the next ATM switch of the wireline ATM network. In other words. an ATM switch is needed to perform upstream-downstream connectivity.

A user terminal can communicate with another within the same cell/cluster as in the case of wireless local area network (LAN) or to an external user terminal through fixed broadband ATM networks [67]. The topology of each WATM cell/cluster can be one of the following types: Broadcast-based ad-hoc WATM. hub-based WATM.

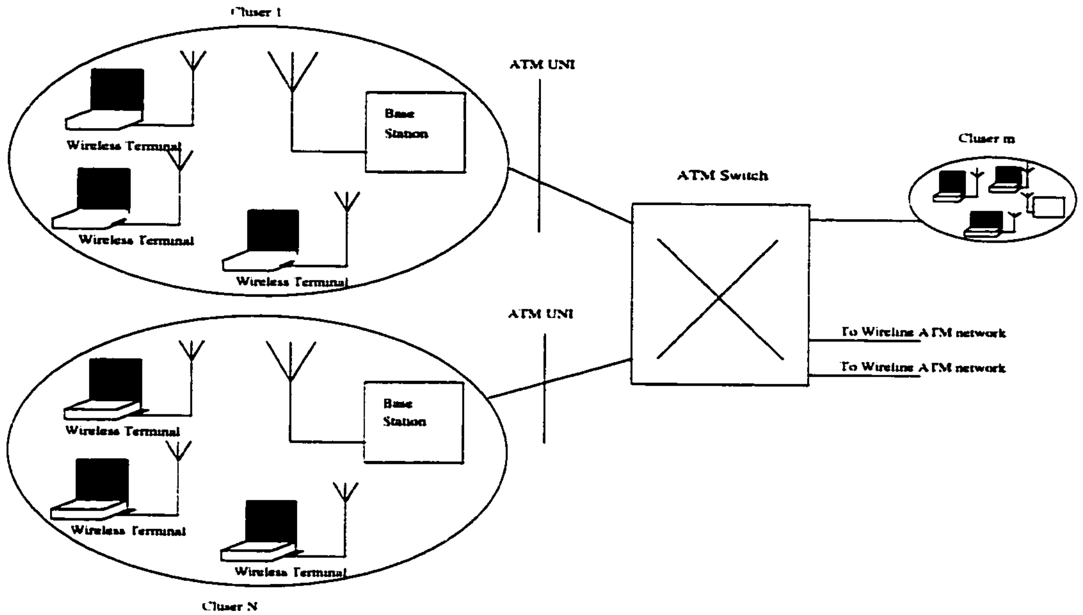

Figure 2.5: The typical architecture of a WATM system.

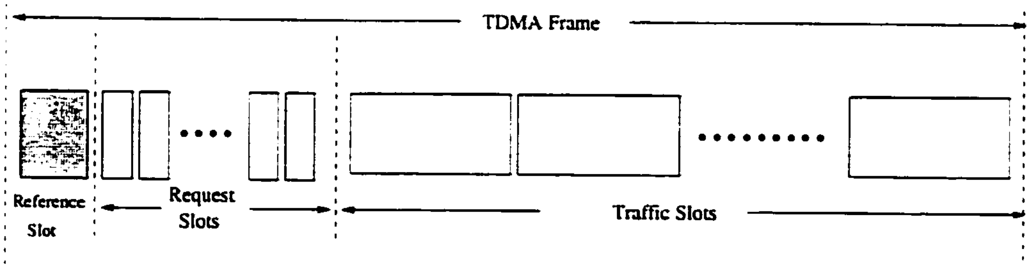

Figure 2.6: The frame structure in a broadcast bandwidth sharing mode (TDMA).

2.2.1 Broadcast based Ad-hoc WATM

In a broadcast-based ad-hoc WATM configuration. a set of N WATM user terminals can share the bandwidth/frame in demand-assignment (DA) TDMA mode. Traffic bursts sent by a terminal contain ATM cells with virtual circuit identifiers (VCI)/virtual path identifiers (VPI). Using the VPI/VCI. the appropriate WATM terminal can receive the WATM cells. The operation of radio link at the physical layer uses only one frequency band as in the dynamic TDMA with TDD MAC protocols.

The user terminals in a cell/cluster can nominate one of them to be the main scheduler and another one as a standby scheduler. At the beginning of each frame, the scheduler can send the frame reference and frame control information indicating the allocation of both request and traffic slots. The request slots can be accessed by the user terminals in a fixed-assignment or slotted-Aloha mode. They are used by the user terminals to send their requests for traffic slots. Following the allocation of the scheduler, the user terminals send their traffic in the assigned traffic slots. If the main scheduler fails, the standby scheduler will take over the control and allocation. Each WATM terminal can transmit during their reserved slots and if not transmitting, can store traffic bursts received from the broadcast medium.

2.2.2 Hub based WATM

In a hub-based configuration, the N WATM terminals send and receive traffic to/from the hub. The hub can be connected to a high-speed ATM backbone network via an ATM switch or multiplexer (Figure 2.7). The hub also acts as the scheduler. A Medium Access Control (MAC) layer is used to control the cell transmission over the radio channel. For the uplink (from user terminals to hub), the MAC protocol can use a dynamic Time-Division Multiple-Access (TDMA) scheme. Since the downlink transmission (from hub to user terminals) operates in a broadcast mode, it can use a Time-Division Multiplexing (TDM) mode. Both uplink and downlink channels can share the same frequency slot as shown in Figure 2.8. In this arrangement, both user terminal and hub operate in a TDD mode. Alternatively, the uplink and downlink channels can occupy a pair of frequency slots and the terminals and hub operate in a full-duplex Frequency-Division Diplexing (FDD) mode as shown in Figure 2.9.

Uplink traffic time-slots are allocated to user terminals on demand by the scheduler (hub in the case of WATM and On-board Scheduler in the case of Satellite ATM (SATM)) using reservation and/or contention in association with the Usage Parameter Control (UPC) mechanism, so that negotiated traffic contract of admitted

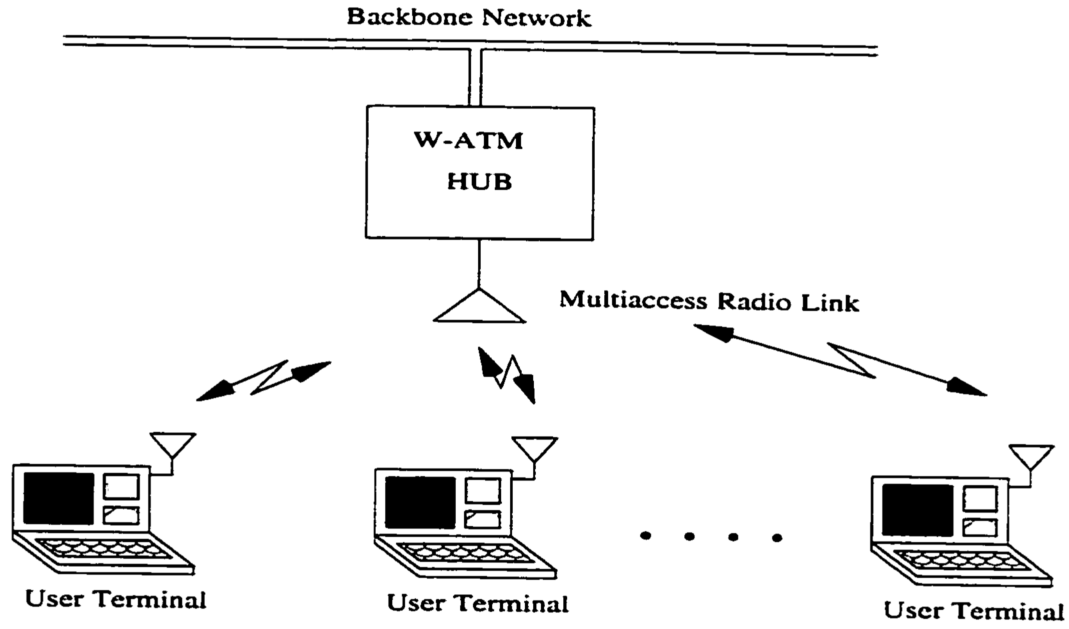

Figure 2.7: A typical example of the hub based cell/cluster operating in a TDMA/TDD mode.

ATM connections can be maintained [2.4]. Each user terminal generally supports 5 categories of multimedia traffic: CBR. rt-VBR. nrt-VBR. ABR. and UBR. Capacity allocation for CBR traffic is straight-forward. Since CBR traffic uses a fixed capacity during an entire connection. its presence affects the system only by reducing the total capacity by an amount of static bandwidth allocated to CBR connections. Other traffic categories, except UBR, are subject to CAC according to the traffic contracts. UBR traffic is given no capacity commitments and, hence, is served by the remaining capacity, if available. ABR traffic can be controlled through feedback, depending on the congestion in the network. On the other hand. rt-VBR and nrt-VBR are uncontrollable within the committed traffic profile. Therefore, we consider the QoS commitments to be satisfied for the uncontrollable rt-VBR and nrt-VBR traffic in the proposed scheduling scheme applicable to both the TDMA/TDD and TDMA/TDM/FDD configurations in the case of WATM and MF-TDMA configuration in the case of SATM.

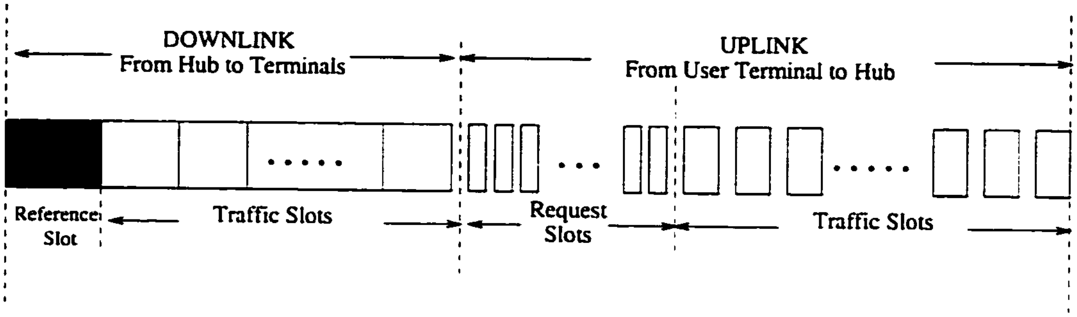

Figure 2.8: The TDMA/TDD frame structure.

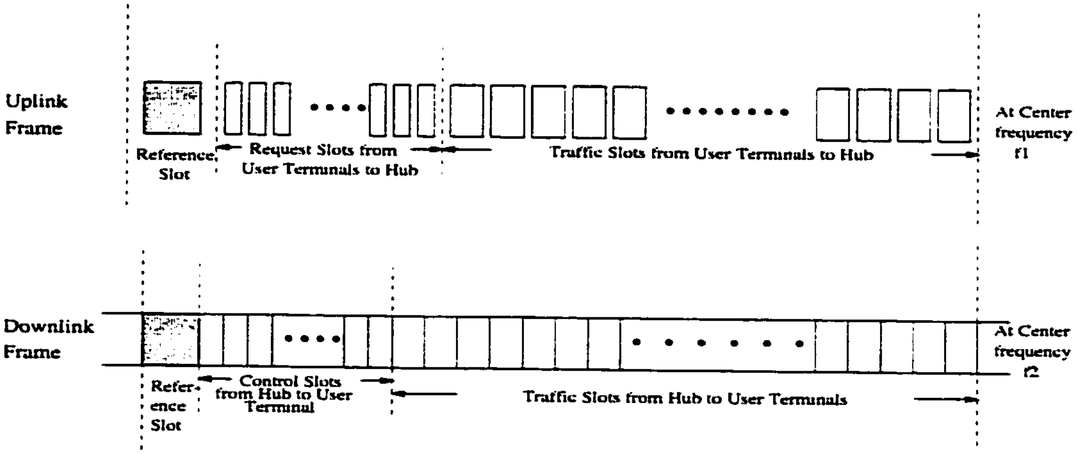
Uplink Frame
Reference Slot
Request Slots from User Terminals to Hub
Traffic Slots from User Terminals to Hub
At Center frequency f1
Downlink Frame
Refer- ence Slot
Control Slots from Hub to User Terminal
Traffic Slots from Hub to User Terminals
At Center frequency f2

Figure 2.9: The TDMA/FDD frame structure.

As shown in Figure 2.8 and Figure 2.9, the uplink TDMA frame contains a frame marker to denote the frame beginning, several signaling time-slots for capacity request and internal control signaling, and a number of traffic time-slots, C. Each non-overlapping traffic time-slot can accommodate one ATM cell. Let C_u represent the uplink capacity. In the case of OBS based satellite ATM, C_u represents the uplink capacity of MF-TDMA. In each frame, the user terminal can send its capacity request in a designated signaling time-slot, containing the numbers of slots needed for rt-VBR and nrt-VBR traffic that has arrived in the previous frame, to the scheduler residing in base-station in the case of WATM and to the scheduler on-board in the case of SATM. Its request can also include other relevant parameters. The scheduler first stores all requests received from different user terminals in a request table. It then calculates the capacity allocated to rt-VBR traffic. If available capacity remains, it continues the capacity allocation to nrt-VBR traffic. After both rt-VBR and nrt-VBR requests are satisfied, if there is still capacity available, the scheduler can give it to all user terminals in a round-robin manner for transmission of ABR and UBR traffic. The scheduler prepares and broadcasts the time-slot assignment to all user terminals, effective in the next frame. In the time-slot assignment, the scheduler need not mention explicitly the number of rt and nrt-VBR packets that the user terminals can transmit. Instead, it only indicates the time-slots allocated to a given user terminal. The user terminal keeps the rt-VBR traffic in a buffer for one frame and the nrt-VBR packets in a FIFO queue since the nrt-VBR traffic can tolerate delay variation. It will transmit the real-time packets first. If the number of rt-VBR packets exceeds the allocated capacity, the excess packets are dropped (lost) as the maximum CTD of rt-VBR traffic must be met. Otherwise, it will continue to send the nrt-VBR packets (stored in its data FIFO queue) in the remaining allocated time-slots. An nrt-VBR cell is lost only when overflow occurs in the queue. Our main objective is to study the effect of the proposed priority-based scheduling on the buffer size of nrt-VBR traffic in order to ensure a specific CLR with the assurance of best possible maximum

CTD and a specific CLR for real-time VBR traffic over the wireless link. Since ATM Traffic Management 4.0 specification (ATM Forum's TM4.0) mentions that there is no need to provide any delay guarantees to non-real-time VBR, we consider only CLR performance of non-real-time traffic. This analysis can be used in designing the buffer size for the nrt-VBR traffic in the user terminal.

In the next chapter, we describe the uplink scheduler structure from the point of view of performance evaluation of RT and NRT traffic. Then, we state the parametric description of the problem statement of the system in order to evaluate the QoS at the uplink and downstream link. In particular, the method proposed for evaluating the uplink scheduling performance in association with traffic characterisation of arrival processes at any downstream link can be used to evaluate the end-to-end performance by using the independence assumption.

Chapter 3

System Modeling

3.1 Modeling of Scheduler Structure

Before we proceed to discuss the modeling of scheduler structure. we present the list of assumptions that we use in our modeling, as follows:

- At the uplink and downlink an ideal channel is assumed which is error free. In using wireless channel when we use very efficient FEC mechanism. the probability of packet loss due to error caused by the channel can be made smaller by which this assumption can be justified.

- The loss or delay due to multiple access mechanism is assumed to be none. This assumption is justified. when the user terminals send their requests for capacity in a dedicated request channel.

- The switch is assumed to be a non-blocking (ideal) switch. This is quite applicable in our case, since in a typical scenario of wireless systems an ATM switch configuration is much smaller as compared to terrestrial switch configuration and therefore. a non-blocking switch is easily affordable.

- We assume the switch to be output queueing based. If we have to consider the general types of input/output queueing based switch, our analysis can be easily modified corresponding to those situations.

- We assume a fixed number of voice, video and data connections in progress from all the user terminals within a beam in the case of Figure 2.1 (correspondingly in the case of Figure 2.5). This quasi-static assumption is quite justified, since by the typical time duration, during which calls arrive and leave, the system reaches a steady state and results in the steady state performance at the packet level.

- We assume symmetric traffic load at each uplink and hence at the output port of the switch. When we relax this assumption we have to calculate the covariance function at different lags presented in Chapter 6, for each uplink which has different amount of load. Thus, our analysis is still can be used even in that case.

- We assume that the packets arriving at each input port from an uplink are probabilistically directed to one of the downlink/downstream link according to a uniform distribution. Relaxing this assumption means that we can have any correlated destination distribution and our analysis can be easily extended even for such cases.

The typical multimedia calls include voice, video and data connections. The queue structure of the user terminals at the uplink beam of Figure 2.1 or at the uplink of Figure 2.5 can be as shown in Figure 3.1. From the viewpoint of evaluating the performance of multimedia connections at the uplink, Figure 3.1 can be simplified as shown in Figure 3.2, since RT packet loss is due to the total number of RT packets exceeding the number of slots in the frame and delay performance of NRT packets at the frame boundaries, is dependent on the NRT packet arrivals from all user terminals.. Each type of connection acts as a variable bit rate input source to the system. This variable bit rate stream is packetised at the user terminal. The aggregated packet generation by a particular type of admitted connections from all the user terminals within a beam in the case of OBS based satellite network in Figure 2.1 or user ter-

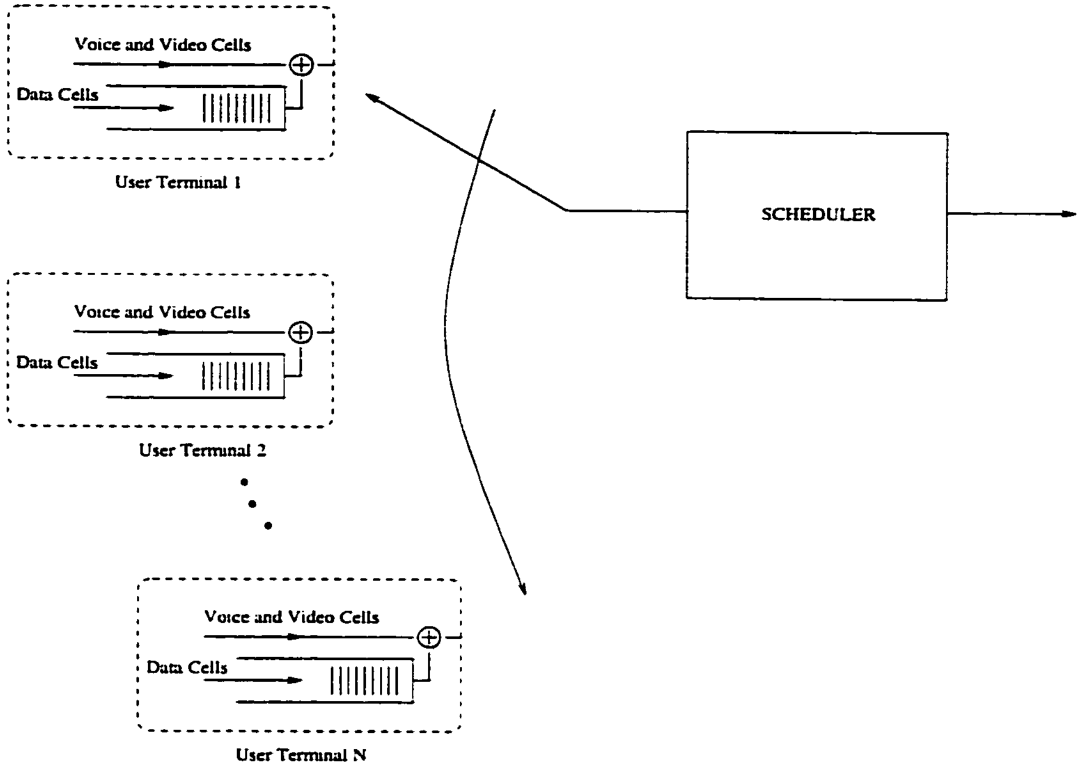

Figure 3.1: The individual queue structure as seen by the uplink scheduler.

minals within a pico/micro/macro cell in the case of WATM of Figure 2.5 can be represented by an appropriate stochastic process. Therefore. the performance will be affected only by the uplink scheduling mechanism of the system. The packets arriving to the on-board packet switch. will be directed to the appropriate downlink and will be queued until the scheduled transmission on the downlink. Similarly, the packets arriving to the scheduler at the base station. will be directed to the downstream link of terrestrial ATM network or to the appropriate downlink of pico/micro/macro cell after being switched by an ATM switch as shown in Figure 2.5. From the point of view of evaluating the QoS, the OBS based satellite system or correspondingly the wireless ATM system can be represented as shown in Figure 3.3 . The generator in Figure 3.3 corresponds to the packet generation process by the stochastic processes corresponding to voice, video and data connections.

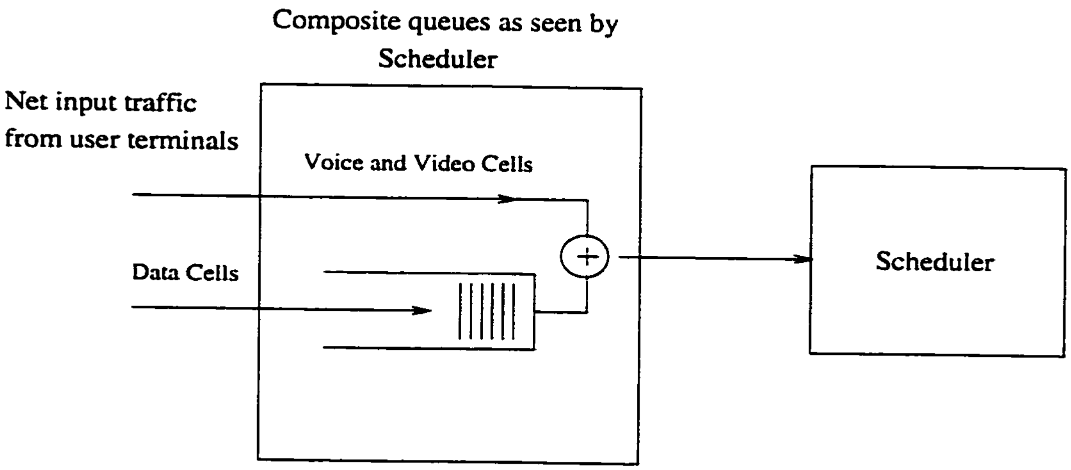
Composite queues as seen by Scheduler
Net input traffic from user terminals
Voice and Video Cells
Data Cells
Scheduler

Figure 3.2: A composite queue structure as seen by the uplink scheduler.

3.2 Problem Statement

Let N_{voi}, N_{vid}, N_{data} be the number of admitted voice. video and data connections respectively at each uplink beam. Each voice and data connection is assumed to be represented by a single On-Off source with exponentially distributed on-times and off-times [58]. During On-periods, an On-Off source generates bits at a fixed rate and does not generate any bit during Off-periods. Each video connection is assumed to be represented by a superposition of a number of independent mini-On-Off sources [57]. The superposed fixed length packet generation process of each voice (N_{voi}), video (N_{vid}) and data (N_{data}) type of traffic, is approximated by a 2-state Markov modulated Poisson process (MMPP), by matching statistical parameters of superposed voice/data/video On-Off sources to 2-state MMPP. The parameters mapped in the matching procedure [68], are the following: mean packet arrival rate. the average load in overload states of superposed On-Off sources (the overload state of superposed On-Off sources are the states in which the rate of the superposed process exceeds capacity), the average load in the underload states of superposed On-Off sources, and the Index of Dispersion for Counts (IDC) at infinite time.

The uplink slots are allocated to the voice. video and data connections at

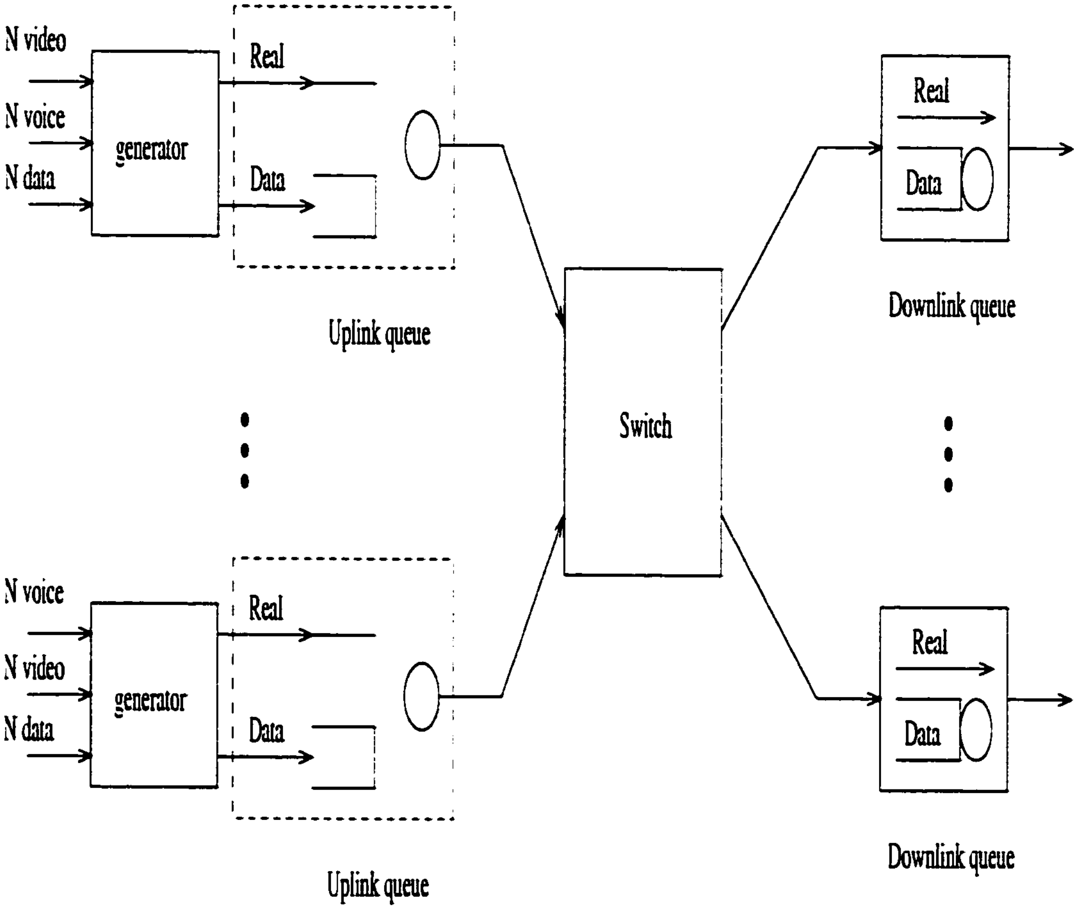

Figure 3.3: The Simplified Model of a Satellite/Terrestrial Wireless Communication System.

the user terminals, by the scheduler. The voice and video (RT) packets are given priority over data packets in transmitting over the uplink. During each frame, the user terminals first transmit the RT packets in the allotted slots. If there are more RT packets than the allotted slots, RT packets which cannot be transmitted during this frame are dropped, otherwise, NRT packets from NRT packet queue will be transmitted for the remaining allotted slots, after the transmission of RT packets. This priority based scheme is followed in order not to have any delay jitter for RT traffic. The packets that arrive at the input port of the switch will be directed to the appropriate output port and queued for transmission at the corresponding downlink/downstream link. The scheduler at the downlink/downstream link uses the same priority mechanism as used at the uplink. The model of the whole system, with Markov modulated Poisson process representing superposition of each traffic type at the uplinks, is shown in Figure 3.4.

The objective is to evaluate the performance in terms of loss for real-time voice and video packets and loss of non-real-time data packets at the uplink, and at the downlink or output port of the on-board packet switch in case of OBS based satellite network and at the uplink and at the downstream link or output port of ATM switch in the case of terrestrial wireless system. The loss probability of NRT traffic corresponding to the finite buffer case is approximated by the probability of the number of NRT packets exceeding that finite buffer size (survivor function) with infinite buffer assumption. Since, the queueing performance with infinite buffer would be much worse than assuming finite buffer (because of bursty traffic still contributing to the queueing performance as compared to the finite buffer case, where they are dropped when the buffer is full), this approximation of survivor function for CLR is a conservative one. The same survivor function can be used to evaluate delay performance of NRT packets at the frame boundaries, which is equal to the number of packets in the queue, multiplied by the slot time. Firstly, modeling of the uplink has to be done, in order to evaluate loss rate of RT packets, loss rate and delay

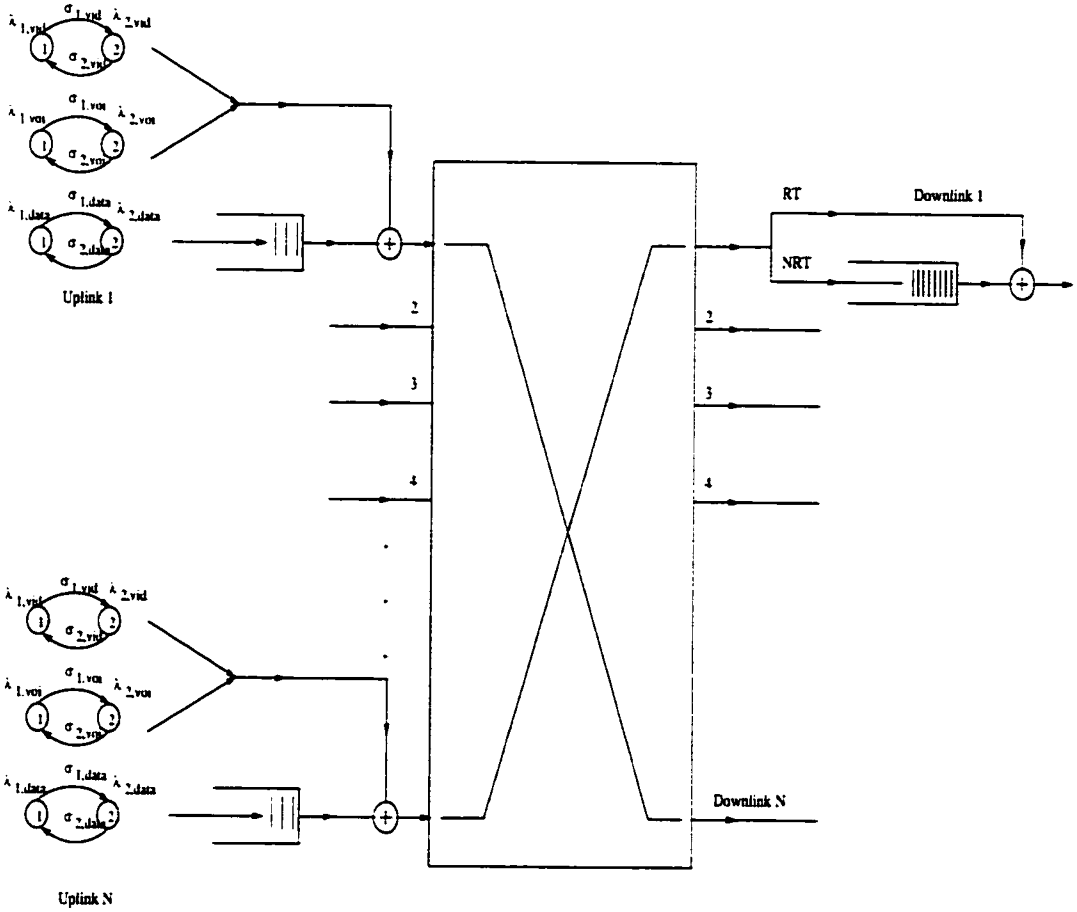

Figure 3.4: The Model of Whole System with Packet Switch.

performance of NRT packets. Secondly, with the assumption of output selection by the RT and NRT traffic arriving from each uplink by the uniform distribution, the modeling of arrival process of RT and NRT traffic has to be done. Thirdly, with the modeling of traffic at the downlink/downstream link, the loss rate of RT packets, loss rate and delay performance of NRT packets have to be evaluated. Using these results, and an independence assumption, if need be, end-to-end performance can be determined.

3.3 Problem Details

3.3.1 Uplink Model

N_{voi}, N_{vid}, N_{data} on-going connections within an uplink can be distributed in any manner among the user terminals. The request for slots for the real-time (RT) voice and video packets and non-real-time (NRT) data packets are received by the scheduler during each frame from the user terminals and it schedules and transmits the scheduling information back to the user terminals. The uplink model, with priority based scheduling mechanism discussed earlier, is as shown in Figure 3.2, where the buffer shown represents the aggregated buffer status of all user terminals within the uplink. We consider this aggregated queueing system, since RT packet loss is due to the total number of RT packets exceeding the number of slots in the frame. Similarly, delay performance of NRT packets is dependent on the NRT packet arrivals from all user terminals. The issue here is the evaluation of RT packet loss probability and loss probability and delay performance of NRT packets at the uplink of Figure 3.2.

3.3.2 Switch Model

An ideal switch is assumed in our case. We assume symmetric traffic load at each uplink and at the output port of the switch. Also, we assume that the packets arriving at each input port from an uplink are probabilistically directed to one of the

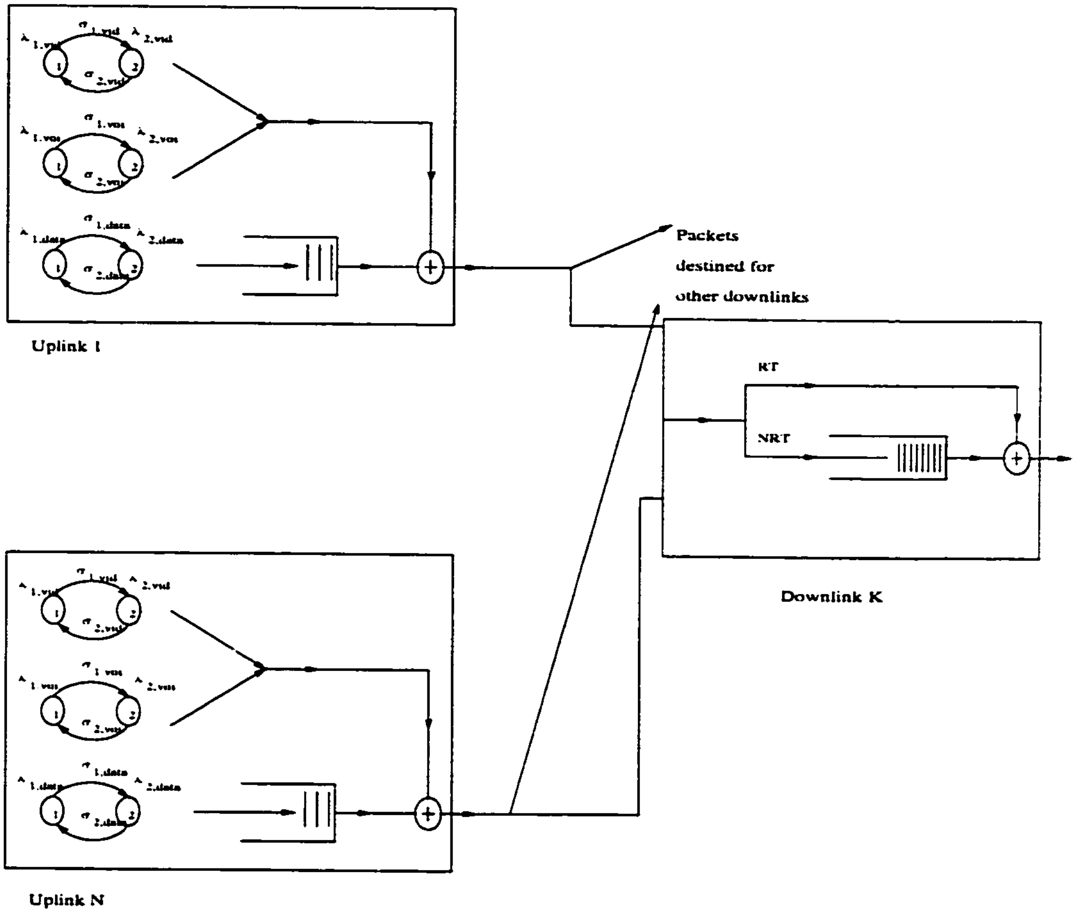

Figure 3.5: The Tandem Link Representation of Uplink and Downlink/Downstream link.

downlink/downstream link according to a uniform distribution. Since we consider only a non-blocking switch no packets are lost inside the switch due to the switching process.

3.3.3 Downlink Model

Because of the Uniform destination assumption and symmetric load assumption, the stochastic nature of the arrival process of real-time packets as well that of the non-real-time packets will be the same at each downlink/downstream link. Thus, the performance of real-time and non-real-time packets at a single downlink is due to the superposed arrival process of packets destined to the downlink, from all uplinks as shown in Figure 3.5. Therefore, the problem here is modeling the arrival process of real-time and non-real-time packets from one uplink to the downlink/downstream link. Later superposition of such arrival processes (from all uplinks) has to be modeled. With the representation of superposed arrival process model for the real-time and non-real-time packets as MMPP's, the loss rate of NRT packets and the loss and delay performance of the RT packets at the downlink/downstream link has to be evaluated. To do this, we use the model developed for the uplink to analyse the performance at the downlink/downstream link. In fact, this methodology can be used to evaluate the performance at any downstream link.

In the following chapter, we discuss various performance evaluation techniques proposed in the literature for solving queueing systems which have similar structure as our analysis leads to. We discuss in detail their methodology: the type of performance measures that can be obtained and the drawbacks associated with such techniques.

Chapter 4

Performance Evaluation Techniques

To evaluate the QoS of the traffic in an OBS based satellite system/Terrestrial wireless system. we can use either simulation or analytical technique. Even though analytical techniques are in general computationally efficient. the analytical modeling used is based on certain assumptions or simplifications of the actual descriptions of the network. As long as the assumptions are justified. the results obtained using analytical techniques will represent closely the actual system behaviour. Thus analytical methods are necessary, to quickly evaluate the performance of the system. Simulation is necessary, to validate the analytical results when the analysis involves approximation. In this chapter, we discuss the OPNET simulation model followed by a discussion on analytical techniques available in the literature.

4.1 Simulation

The simulation model is developed using OPNET. OPNET models are hierarchical to naturally parallel the structure of actual communication networks. OPNET allows protocol level simulation also. However. for our modeling, we need mainly the abstract queueing and switching phenomenon to be implemented. OPNET uses three hierarchical modeling domains called Network, Node. and Process domains. The network level modeling corresponds to interaction among multiple networks. However, since we have a single OBS based satellite network system(correspondingly a single ATM switch for terrestrial wireless system), the simulation model had only one Network model. This network model used a single node model. As mentioned in Chapter 3, we developed a simulation model for the aggregated queueing system. Therefore, we have a single generator at each uplink with fixed number of voice, video and data connections represented by MMPPs, that generate RT and NRT packets for the aggregated system. We consider a 4 by 4 system and the simulation model is developed corresponding to Figure 3.4. The description of the node and process model are given in Appendix A.

The confidence interval for occupancy probability from a single simulation is calculated by using the method of batch means. The number of transient samples to be discarded is determined by running simulation independently and plotting the distribution function of each random variable. When the distribution function seems to converge to the same empirical distribution, all the sequences previous to this point are considered transient. We denote this number by n_0. Thus, in a single simulation we discard the samples starting from 1 to n_0. Next, in order to determine the batch size, we have to make sure that the correlation between random variables of queue occupancy from one batch to the contiguous batch is negligible. For this purpose, we estimate $\hat{\rho}(k)$, where $\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$. Let X_n be the sample value of the sequence. Thus, $\hat{\gamma}(k) = \frac{1}{N}\sum_{n=1}^{N-1}(X_n - \bar{X})(X_{n+k} - \bar{X})$, where $\bar{X} = \frac{1}{N}\sum_{n=1}^{N} X_n$. If for $|k| > L$, $\rho(k) = 0$(the correlation function $\rho(k)$ is correlation between two members of the sequence with a separation of k units), then most of $\hat{\rho}(k)$(approximately 95%) should lie between the limits $\pm\mathbf{D}$ where $\mathbf{D} = 2\left(\sum_{k=-L}^{k=L} \frac{\hat{\rho}^2(k)}{N}\right)^{1/2}$[55]. The value of L is chosen as equal to that value of k beyond which $\hat{\rho}(k)$ lies within $\mathbf{D}$. Then each batch size N is chosen as $N = J * L$. Let $N' + n_0$ be the number of samples generated in a single simulation run, such that $M = \frac{N'}{N}$ represents M different batches. Thus we have M batches of sequences. Let X_{mn} denote the n-th sequence value of batch m.

Let $\hat{p}_{im} = \frac{v_{im}}{N}$, where v_{im} is the number of X_{mn} in batch m, taking values equal to i. Therefore, $E(\hat{p}_{im}) = p_i$. Let the estimator of probability of queue occupancy being equal to i be $\hat{p}_i$. Thus $\hat{p}_i = \frac{1}{M}\sum_{m=1}^{M}\hat{p}_{im}$ and $s^2(\hat{p}_{im}) = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{p}_{im} - \hat{p}_i)^2$. The random variable $\frac{\hat{p}_i - p_i}{s(\hat{p}_{im})/M^{1/2}}$ has approximately t-distribution with $M-1$ degrees of freedom. The confidence interval is given by

$$Pr\{\hat{p}_i - t_{M-1}(1-\frac{\alpha}{2})s(\hat{p}_{im})/M^{1/2} \leq p_i \leq \hat{p}_i + t_{M-1}(1-\frac{\alpha}{2})s(\hat{p}_{im})/M^{1/2}\} \approx 1-\alpha$$

where $t_n(x)$ is the $100x$-th percentile of t-distribution with n-degrees of freedom.

The advantage of developing such a simulation model is that it represents the actual description of the network. However, the main disadvantage is that it takes lot of time (of the order of days) to yield dependable results, in particular for the evaluation of very low probability of packet losses.

4.2 Candidate Analytical techniques

We consider the analytical techniques available in the literature to solve for queueing performance at a single multiplexer. The packets generated by the multimedia traffic are queued for transmission at the multiplexer. The scheduling discipline at the multiplexer can be FIFO(First in First out), or priority based for accessing the output link or the buffer. The general queueing equation of the buffer at the multiplexer can be written as,

$$q(\tau_n) \;=\; \left(q(\tau_{n-1}) + \sum_{k=1}^{N} A_k(\tau_n, \tau_{n-1}) - C_n\right)^{+} \qquad (4.1)$$

where τ_n is the n-th epoch at which queueing system is observed and $A_k(\tau_n, \tau_{n-1})$ is the random variable representing the number of arrivals of type k traffic at the buffer during the interval (τ_n, τ_{n-1}) and C_n is a random variable representing the number of units from the buffer that can be served during this interval. The queueing system is assumed to be stationary.

For the case of ATM with fixed size packets called cells, τ_i's can be the beginning or end of the i-th slot, where each slot time(say Δ) corresponds to transmission of one ATM cell over the output link and C_n can be equal to some constant C(usually it is taken as 1)[35]. Letting $\tau_n = t + \Delta$ and $\tau_{n-1} = t$, Equation 4.1 can be rewritten as,

$$q(t+\Delta) = \left(q(t) + \sum_{k=1}^{N} A_k(t, t+\Delta) - C\right)^{+} \tag{4.2}$$

For the case of fixed size frame based transfer, τ_i's are time epochs of frame start or end and C is the number of slots in a frame. Equation 4.1 also describes other models. For example, for the case of Markovian arrival process based queueing systems, τ_i's are departure epochs of packets and C_n is equal to 1. The interval (τ_i, τ_{i-1}) when the system is not empty is generally distributed, since it corresponds to the service time of a packet and $A_k(\tau_i, \tau_{i-1})$'s correspond to arrivals due to a Markovian arrival process.

In each case, the buffer size can be finite or infinite. To evaluate packet loss probabilities for a finite buffer, the tail probability corresponding to that finite buffer size from the infinite buffer case has been used as an effective approximation for large buffer sizes [33]. When the service process can be modeled similarly to the arrival process, the service process together with the arrival process can be considered as the modified arrival process as far as evaluating the performance of the system is concerned. Since the capacity of the system is finite, modeling the variable service process results in a finite state stochastic process. Therefore, effectively, a queueing process with constant server type can be used[9][44][47]. In general, there are three types of techniques used in solving queueing problems. They are matrix analytic, generating function or transform based, and fluid flow approximation based techniques. First, we present the details of the matrix-analytic technique. Following this, we discuss the technique based on the generating function method. Finally, we discuss the performance evaluation technique based on fluid flow approximation.

4.2.1 Matrix-Geometric Technique

When the infinite matrix representing the transition matrix has the special structure of M/G/1 type or G/M/1 type, then we can use the matrix-analytic technique outlined in [37]. The theme of the discussion is based on first passage distribution by one level(a level is a transition from a state, say (i, k), where the first index corresponds to the number of packets in the system with $i \in \{0,1, \dots \infty\}$ and the second index corresponds to the phase of the underlying modulating process with $k \in \{1, 2, \dots m\}$ where m is the maximum number of phase states, to the state $((i-1), l)$ due to departure in case of Ph/GI/1(Ph represents phase-type distribution) queue. Correspondingly in the case of GI/Ph/1 queue it is the transition from the state, (i, k) to the state $((i+1), l)$ due to an arrival) in a Markov renewal process which is skip-free to left or skip-free to right. The skip-free to left property for Ph/GI/1 queue is due to the structure of the matrix, which exhibits the property that reaching level **0** from level **i** can only be done by visiting each level in between. We discuss briefly the matrix-geometric algorithm for the case of BMAP/G/1 queue, in this section.

The Batch Markovian Arrival Process(BMAP) is a two-dimensional Markov process (N(t), J(t)) on the state space $\{(i, j) : i \geq 0, 1 \leq j \geq m\}$ with an infinitesimal generator Q having the structure,

$$Q = \begin{pmatrix} D_0 & D_1 & D_2 & \dots \\ 0 & D_0 & D_1 & \dots \\ 0 & 0 & D_0 & \dots \\ \cdot & \cdot & \cdot & \dots \\ \cdot & \cdot & \cdot & \dots \end{pmatrix}$$

where D_k, $k \geq 0$, are $m \times m$ matrices. D_0 has negative diagonal elements and represents an auxiliary state or phase variable then the above Markov process defines a batch arrival process where a transitions from a state (i, j) to state $((i+k), l)$, k≥1, 1≤j, l≥m, correspond to batch arrivals of size k, and thus, the batch size can depend on i and j.

Let the service times have an arbitrary distribution function, $\tilde{H}(x)$, with Laplace-Stieltjes transform, $H(s)$ and finite mean μ_1'. The BMAP and service process are assumed to be independent, with traffic intensity, $\rho = \frac{\lambda_1'}{\mu_1'} < 1$, where λ_1' is the mean arrival rate. The embedded Markov renewal process at departure epochs is defined as follows. Let $(\tau_n : n \geq 0)$ denote the successive epochs of departure (with $\tau_0 = 0$) and let us define X_n and J_n to be the number of customers in the system and the state of BMAP at τ_n^+. The sequence $(X_n, J_n, \tau_{n+1} - \tau_n)$ forms a semi-Markov process. This semi-Markov process is positive recurrent when the traffic intensity is less than 1. The transition probability matrix is given by,

$$\tilde{P}(x) = \begin{pmatrix} \tilde{B}_0(x) & \tilde{B}_1(x) & \tilde{B}_2(x) & \dots \\ \tilde{A}_0(x) & \tilde{A}_1(x) & \tilde{A}_2(x) & \dots \\ 0 & \tilde{A}_0(x) & \tilde{A}_1(x) & \dots \\ 0 & 0 & \tilde{A}_0(x) & \dots \\ . & . & . & \dots \\ . & . & . & \dots \end{pmatrix}$$

for $x \geq 0$. For $n \geq 0$, $\tilde{A}_n(x)$ and $\tilde{B}_n(x)$ are the m × m matrices of mass function defined by

$[\tilde{A}_n(x)]_{ij} = Pr\{$Given a departure at time 0, which left at least one customer in the system and the arrival process is in state i, the next departure occurs no later than time x with the arrival process in state j, and during that service there were n arrivals$\}$,

$[\tilde{B}_n(x)]_{ij} = Pr\{$Given a departure at time 0, which left the system empty and the arrival process is in state i, the next departure occurs no later than time x with the arrival process in state j, and during that service there were n arrivals$\}$.

Queues with embedded Markov renewal processes whose transition probability has the structure just shown above are referred to as queues of the "M/G/1 type" or queues of the "M/G/1 paradigm". Let us define $\tilde{G}_{jl}^{[1]}(k; x)$ for $k \geq 1$ and $x \geq 0$, to be the probability that first passage from state $((i+1), j)$ to the state (i, l), where $i \geq 1$, and $1 \leq j, l \leq m$, occurs in exactly k transitions and no later than time x. Let

$\tilde{G}^{[1]}(k;x)$ be the matrix with elements $\tilde{G}^{[1]}_{jl}(k;x)$. Let

$$G(z,s) = \sum_{k=1}^{\infty} \int_0^{\infty} e^{-sx} d\tilde{G}^{[1]}(k;x) z^k$$

By using the first passage argument, it has been shown that $G(z,s)$ satisfies the following non-linear matrix equation [80],

$$G(z,s) = z \sum_{v=0}^{\infty} A_v(s) G^v(z,s)$$

where $A_v(s)$ is the Laplace transform matrix of $\tilde{A}_v(x)$. By denoting, $G(1,0) = G$ and $A_v(0) = A_v$, we can write,

$$G = \sum_{v=0}^{\infty} A_v G^v$$

It has been shown that $G(z,s)$[87] also satisfies, in the case of BMAP/G/1 case,

$$G(z,s) = z \int_0^{\infty} e^{-sx} e^{D(G(z,s))x} d\tilde{H}(x)$$

where $D(G(z,s)) = \sum_{j=0}^{\infty} D_j G^j(z,s)$. Therefore, the matrix G is the root of $G = \int_0^{\infty} e^{D[G]x} d\tilde{H}(x)$. This particular relationship provides an efficient algorithm for computing the matrix G, with the use of uniformisation techniques. The steady-state vector of G satisfies $\mathbf{g}G = \mathbf{g}$ and $\mathbf{ge} = 1$. The probability vector $\mathbf{x}_0$ can be obtained by considering the classical property of Markov chains, the mean recurrence time of the state $(0,j)$ in the Markov chain $\mathbf{P}$, where $\mathbf{P} = \tilde{P}(\infty)$[87]. Finally we can use recursion to solve $\mathbf{x}_i$ where $\mathbf{x}_i$ represents the steady state probability vector of i packets in the system and the BMAP being in various phases.

The more general analysis for the case of BMAP/SMSP(SMVP)/1 queue has been presented in [39]. In this analysis, the embedded Markov analysis has an additional dimension as compared to the above mentioned analysis of BMAP/G/1 queue, due to the Semi-Markov Service Process. In the following section, we discuss the z-transform technique to solve for the case of queues with infinite queue size at the multiplexer.

4.2.2 Solution Technique using Probability Generating Function Methods

For the systems with infinite queue, when the system of equations, either all or all except a few boundary equations, can be represented as a difference equation, then we can use the probability generating function method. From the difference equation, we can write down the probability generating function. This probability generating function usually will have unknown boundary probabilities which can be solved for, by solving a system of linear equations obtained from the numerator of the generating function. To construct the system of simultaneous linear equations, we have to solve for roots of the denominator of the probability generating function. This task becomes numerically intractable as the size of the system grows. Typical examples of probability generating function derivations with Poisson arrival process and embedded Markov chain analysis for queues can be found in [38] and [40].

With multimedia traffic, the correlated arrivals have to be modeled as a modulated arrival process. This can be done by approximating the traffic by discrete-time Markov chains. The queueing process becomes a two-dimensional stochastic process corresponding to the queue size and state of the arrival process. For the case of discrete-time analysis of voice and data integration at the multiplexer, the data queuing performance with and without speech activity detectors for voice traffic is analysed in [41]. In their analysis, voice packets which cannot be transmitted in a frame were considered to be dropped and therefore only the data packets were queued. in an infinite queue. A relatively different model with the queueing of voice packets along with data packets in an infinite queue was considered by [42]. The boundary probabilities were obtained by finding the roots of the characteristic equation of the probability generating function, in both cases[41][42]. It is mentioned in [42](page no. 1267), that the numerical analysis may be intractable with large link capacities. To overcome this difficulty, an alternate method of finding the roots using the spectral decomposition technique, was considered in [43].

We discuss the generating function approach, which is based on the spectral decomposition technique, to analyse the discrete-time queues with arrival process modeled as a Markov chain[43]. The importance of this approach lies in the fact that it can be easily extended to the case of arrival and service processes modeled as superposition of multiple Markov chains[43][44]. This technique can also be applied to the case of fluid flow modeling involving arrival process as a superposition of multiple Markov processes[43].

It is considered, in [43] the queueing system in discrete-time domain representing a TDM system, and evolving as according to the following equation (refer to Equation 4.1),

$$q(n+1) \quad = \quad \left(q(n) + \sum_{k=1}^{N} A_k(n, n+1) - M\right)^{+} \tag{4.3}$$

where $q(n)$ is the queue size at the beginning of n-th interval. An interval is the basic unit of time used for the discrete-time queueing analysis. For example, it can be the frame time interval, where each frame contains M slots and during each slot a packet of fixed size can be transmitted by the output link.

In the following discussion we assume there is only one type of input traffic in the system. Let the arrival process be modeled as a one-dimensional homogeneous Markov chain, with the transition probability matrix P whose elements $p_{m,n}$ stand for $p_{m,n} = P\left(A(n+1) = n / A(n) = m\right)$ for $0 \leq m, n \leq N$, where $N+1$ is the total number of states of the arrival process.

Thus, the Markov chain representing the queueing process is a two-dimensional Markov chain. Let its steady state joint probability of $q = i$ and $A = j$ be written as $p_{ij} = P\left(q = i, A = j\right)$. These steady state probabilities satisfy the following equation as shown below.

$$p_{ij} \quad = \quad \sum_{l=0}^{min(N,M+i)} \sum_{k=0}^{M-l+i} p_{kl,ij} p_{kl} \tag{4.4}$$

with $p_{kl,ij} = \begin{cases} p_{l,j} & if \quad i = (k+l-M)^{+} \\ 0 & \quad otherwise \end{cases}$.

Let us define the generating function

$$Q_j(z) \;=\; \sum_{i=0}^{\infty} p_{ij} z^i \qquad (4.5)$$

for every $0 \leq j \leq N$ with $E(z^q) = \sum_{j=0}^{N} Q_j(z)$. Let $Q(z) = [Q_0(z), Q_1(z), ...Q_N(z)]^T$. Then from Equations 4.3, 4.4 and 4.5, we can write down the following expression for $Q(z)$[43].

$$Q(z) \;=\; \left[I - z^{-M} P^T diag\{z^i\}\right]^{-1} P^T diag\{z^i\} B(z) \qquad (4.6)$$

where $B(z) = [B_0(z), B_1(z), ...B_N(z)]^T$, with

$$B_j(z) = \begin{cases} \sum_{k=0}^{M-l-1} p_{kl}(z^{-l} - z^{k-M}) & if \;\; l < M \\ 0 & otherwise \end{cases}$$

The vector $B(z)$ involves the unknown boundary probability terms p_{kl} for $0 \leq k+l < M$. To solve for these boundary probabilities, the technique proposed in [43] is as follows.

First the matrix $P^T diag\{z^i\}$ is diagonalised. Thus.

$$P^T diag\{z^i\} = G(z)\Lambda(z)G^{-1}(z)$$

where $\Lambda(z)$ is a diagonal eigenvalue matrix of $P^T diag\{z^i\}$, given by $\Lambda(z) = diag[\lambda_0(z), \lambda_1(z), ..., \lambda_N(z)]$. For each given $\lambda_j(z)$, let $g_j(z)$ and $h_j(z)$ be the respective left column and right row eigenvectors. Therefore. by spectral decomposition.

$$P^T diag\{z^i\} = \sum_{j=0}^{N} \lambda_j(z) g_j(z) h_j(z)$$

By using this equation. $Q(z)$ can be simplified to.

$$Q(z) = \sum_{i=0}^{N} \frac{z^M \lambda_i(z)}{z^M - \lambda_i(z)} g_i(z) h_i(z) B(z)$$

Since. $g_i(z)$ and $h_i(z)$. for every i are the eigenvectors. they must be analytic for $abs(z) \leq 1$. Let $\Pi_i[z^M - \lambda_i(z)]$ be the system characteristic function. The poles are equal to the roots of this function.

For each of these roots, we can establish a boundary equation. All the boundary probability values are then obtained by solving these linear boundary equations along with the equation of the first derivative of $Q(z)$ at $z = 1$. Once we can express $Q(z)$ in terms of these boundary probabilities, we can obtain moments of queue length. The details of this method and the extension to superposition of multiple independent discrete-time Markov chain subprocesses are described in [43]. In [44], these details are further extended to express the queue length tail distributions by using the roots of the system characteristic function.

In [45] a set of homogeneous correlated on-off discrete-time sources with number of packets generated being expressed by general distribution(with number of packets > 1), was assumed to represent the arrival process for the ATM multiplexer. An infinite queue length for the multiplexer was assumed. A functional equation representing the joint probability generating function(PGF) of number of fixed size ATM packets in the queue and the number of sources in the on-state. could not be expressed in a closed form in [45]. But, in [46] a closed form expression was developed for the PGF of joint distribution of number of packets in the queue and the number of sources in the on-state at slot k starting from slot 0 with both zero and non-zero initial conditions. This closed form expression was developed by using induction[46]. Using the Final Value theorem, a closed form expression for the PGF of joint distribution could finally be obtained from this transient analysis[46]. Also in[46], the analysis was further extended to the case of heterogeneous on-off sources.

4.2.3 Fluid Flow Model

The above analyses were based on the assumption of packet arrival process is of the type of generalised Markovian arrival process. However, the arrival process of voice traffic[75] is of the on-off type, with during on-times, the source outputs at a constant bit rate into the multiplexer. Even though the video traffic is of variable bit rate source, its statistical characteristics can again be approximated by a superposition of a number of On-Off sources[57]. The bit stream output from On-Off source will be packetised and sent into the network. The above analyses, therefore, take into consideration the packetised arrival process. However, it is also possible to analyse the multiplexer queue using fluid flow model with On-Off sources, which approximates the multiplexer system with packet queue. Consider the queueing equation from Equation 4.2. We can take the fluid approximation as $\Delta \to 0$ and thus, the same problem can be formulated in the continuous time domain, as

$$\begin{aligned} q'(t) &= lim_{\Delta\to 0}\frac{q(t+\Delta)-q(t)}{\Delta} \\ &= lim_{\Delta\to 0}\left(\sum_{k=1}^{N}\Delta^{-1}A_k(t,t+\Delta)-C\Delta^{-1}\right) \end{aligned} \tag{4.7}$$

with q(t) ≥ 0. $lim_{\Delta\to 0}\Delta^{-1}A_k(t)$ represents the arrival rate of the k-th type at time t, which is modeled as a Markov process and $C\Delta^{-1}$ represents the link transmission rate. Also $q(t)$ is characterised as a continuous variable and can be solved by using the methodology outlined in [56].

The essential approximation involved with this modeling is the assumption of continuous variation of the buffer size ignoring the variation due to discrete packet size. This assumption is valid only when the number of sources to be multiplexed and the capacity of the link are large, compared to the discrete buffer size, because of cell arrivals and departures. We discuss briefly the method involved in deriving the survivor function of the queue occupancy(the survivor function of a particular queue occupancy value is defined as the probability that the queue occupancy exceeds that value). Let N_1 be the number of On-Off sources of type 1 with constant bit rate of A_1 when the sources are in "On" state and N_2 be the number of On-Off sources of type 2 with constant bit rate of A_2 when the sources are in the "On" state. Let a and b be the transition rate of going from "Off" state to "On" state and "On" state to "Off" state of type 1 sources. Similarly, let c and d be the corresponding transition rate of type 2 sources. The distribution of sojourn time in all states is assumed to be exponentially distributed. By considering the Chapman-Kolmogrov forward equations for the joint probability distribution of source state and multiplexer queue size, at steady state we can write down a set of simultaneous differential equations for $F_{ij}(y) = Prob\{$input source is in state (i, j) , multiplexer queue size $\leq y\}$ for $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$. The solution of the vector $\mathbf{F}$ with components $F_{ij}(y)$ is expressed in terms of the eigenvalues of the matrices involved, when the simultaneous system is written in a matrix form. The method involved in obtaining these eigenvalues and eigenvectors, in order to obtain the survivor function of the multiplexer queue, is outlined in [59].

The matrix analytic technique usually has better numerical stability compared to the transform based technique due to its algorithmic nature. Also, with this technique, there is no need for transform domain matrix manipulations and transform inversions. However, the matrix analytic techniques use iterative algorithms to find the minimal nonnegative solution of certain nonlinear matrix equations arising in the Markov chains. The low linear convergence rates of iterative algorithms employed makes use of this method not viable for large dimensionality, especially under heavy traffic load. The transform technique has to determine the zeros of the determinant of a certain polynomial matrix within the unit disk. These zeros are used to obtain the set of linearly independent equations, the solution of which gives the desired unknown boundary probabilities. Also, to obtain unknown steady state probabilities, the transform inversion has to be carried out. This approach has problems due to numerical properties. The fluid flow approximation leads to large errors in predicting the buffer occupancy when the buffer size is small, i.e., the cell level loss behaviour[69].

Because of the difficulties associated with the transform technique and the fluid flow approximation technique, we use the matrix analytic technique in our analysis. Our proposed model uses On-Off source models and approximates the superposition by Markov Modulated Poisson Processes. Our queueing analysis leads to M/G/1 type structure. In the following chapter, we outline the nature of the input traffic and its representation by 2-state MMPPs. Following this, we describe the priority based scheduling strategy and its analytical representation. Also, we present an embedded Markov chain analysis, to evaluate the performance of RT and NRT traffic, by using the matrix-geometric technique.

Chapter 5

Performance Evaluation of Priority Based Scheduling at the Uplink

In assigning the slots. the scheduler gives priority to the real-time traffic. as the end-to-end delay requirement of real-time traffic has to be met. If the real-time traffic demand cannot be satisfied during the current frame. those real-time fixed size packets that cannot be transmitted are dropped rather than being queued for later transmission. This policy is adopted in order that real-time traffic has minimum possible delay and delay jitter. The analysis can be easily extended to the case where the dropping is done when a finite buffer of size greater than the total number of slots in a frame becomes full. On the other hand. the non-real-time fixed size packets are queued, and the status of the individual queues at the user terminal is available in the request update table at the scheduler.

5.1 Analytical Modeling of Uplink

We analyse the performance of the uplink multiplexing system. This analysis will be used in evaluating the performance of the downlink once the arrival process to the downlink is characterised. We assume a fixed number of C slots being devoted to the uplink transmission. The main focus of our analysis is to evaluate the cell loss ratio of real-time packets and queue length distribution of non-real-time packets due to the priority scheduling, as seen by the scheduler. We have to bear in mind that the non-real-time queue of Figure 3.2 represents the sum of the individual queues at the user terminals of Figure 3.1. We can design the buffer size of user terminal for non-real-time traffic to be the buffer size of Figure 3.2 needed to satisfy a particular CLR value.

5.1.1 Traffic Models

Within the capacity constraints imposed by the channel, there is no restriction on the type of traffic that can be supported. There are different traffic models for characterising different types of applications. Each of these modeling techniques matches certain statistical characteristics of traffic to the approximating model. Appropriate traffic models have to be chosen so that queueing analysis of scheduling strategy, is tractable. For example, the superposition of video sources is modeled as a continuous-time Markov process[57]. While it may be true that a continuous state autoregressive model may be more accurate, it is not amenable to analysis by queueing theoretic techniques[57][1].

The multimedia traffic to be supported is categorized into three broad classes: 1) voice, relatively low bandwidth real-time traffic but in large volume, 2) video, wideband real-time traffic and 3) data, non-real-time traffic. Typical examples of non-real-time data traffic are remote data-base access, client-server interaction and http access.

It is well known that the characteristics of voice sources can be well approximated by on-off sources [75]. Also the characteristics of aggregate video sources of video-telephone connections can be well represented as a superposition of mini-on-off sources[57]. Of course, there are widely varying models based on multi-state Markov chain and autoregressive models[49][48][50], depending on the nature of the video type. Most of the non-real-time data applications can be approximately char-

[1] See page no. 837 of [57] for further discussion.

acterised by on-off models[85]. Data traffic from LAN gateways and some Internet Protocol(IP) traffic can be long-range dependent. However, it can be modeled by appropriate Markovian arrival process (MAP) models as shown in [51]. Since, our modeling technique can be used for any MAP representation, both real-time traffic and non-real-time traffic can be represented by any MAP. However, for simplicity, we consider the superposition of real-time voice, real-time video and non-real-time data sources, each to be represented by 2-state MMPP. Since the queueing analysis will be much more complicated with on-off source representation, as the number of sources increases (due to the explosive growth of state space), we rather use the approach of modeling the superposition of homogeneous on-off sources by a 2-state MMPP model[84]. We use a procedure outlined in [68] which is similar to [84] and [69]. In this procedure each video source is assumed to be represented by a superposition of homogeneous mini-on-off sources[57](say m). Therefore, a superposition of N video sources will be represented by a superposition of $m * N$ homogeneous mini-on-off sources[52]. The next step needed for our queueing analysis is the approximation of superposition of homogeneous on-off sources by 2-state MMPP. Even though the technique initially proposed in [84] for this purpose could be used to evaluate the multiplexer performance, the approximate cell loss probability was too far from reality. Therefore, there were different techniques proposed in the literature to model the superposition of on-off sources by Markov modulated processes[69][53]. In [54], the authors propose a technique based on the characterisation of the aggregate cell arrival rate and variance-time curve, to model the superposition of heterogeneous on-off sources, by 2-state MMPP. In [97], the technique based on the characterisation of probability mass function of number of cell arrivals in an observation interval, index of dispersion for counts(IDC) values at time t and at infinite time and average cell arrival rate, has been proposed to model multimedia traffic, by the superposition of N identical 2-state MMPPs. From the work in [97], we can infer that, since the probability mass function of number of arrivals in a packet time is binomially distributed.

for the case of superposition of homogeneous sources, a single 2-state MMPP can be used. Thus, the results we present in this paper are restricted to only for sources that are representable by on-off sources with exponentially on times and off times or by 2-state MMPP, although the modeling can be easily extended for any MAP and hence for any traffic source representable by MAP.

5.2 Queueing Model of the Scheduler

Since real-time traffic has stringent requirements on end-to-end delay, the scheduler gives priority over non-real-time traffic in performing dynamic allocation of MF-TDMA slots to the satellite user terminals and similarly on uplink TDMA slots to the WATM user terminals. Let C be the total number of traffic time-slots in each uplink TDMA frame. The scheduling of cell transmissions can occur only at the beginning of each frame time. If the sum of real-time cell arrivals during a frame time at each user terminal exceeds C(where C is the number of slots in a MF-TDMA frame in the case of OBS based satellite network and the number of slots dedicated on the uplink TDMA in the case of WATM network), then the excess is lost. The non-real-time data packets arriving during a frame period are queued at the terminals. Thus, at the beginning of each frame, if there are real-time packets that have arrived during the previous frame period, they all will be transmitted first and if the number of real-time packets is less than C, then the slack is filled by non-real-time data packets from the data queue of each station, as assigned by the scheduler. Note that by queueing the real-time packets, the real-time cell loss can be reduced. However, this will introduce delay and jitter for real-time traffic. Thus, the queueing model of the scheduler can be represented as shown in Figure 5.1.

Because of the scheduling policy, the real-time packets suffer loss and their delay is bounded at the uplink by one frame duration plus the propagation delay. Since non-real-time traffic is sensitive to loss, the CLR has to be controlled when

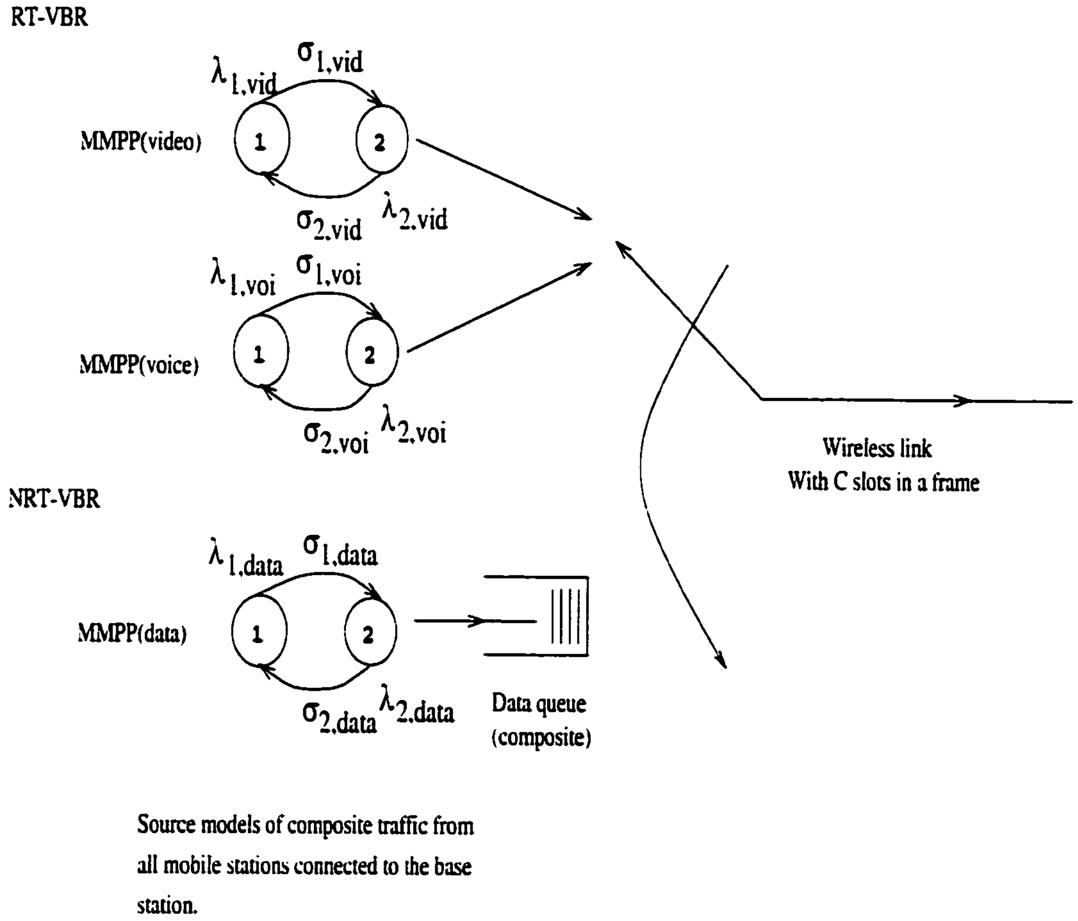

Figure 5.1: The queueing model for real-time and non-real-time traffic at the uplink of a beam.

finite buffers are allocated to non-real-time traffic. In our analysis, we assume that the buffer size for non-real-time traffic is infinite. For finite buffer cases, we can approximate the CLR by the tail probabilities of the infinite buffer case. The non-real-time traffic queue considered actually represents the sum of the non-real-time traffic queues at the terminals within an uplink. Therefore, the finite queue size corresponding to a particular CLR can be set at each user terminal(even though it represents the worst case scenario). The survivor function of queue length is defined as the probability that the queue length exceeds a certain value, in other words it is the complementary distribution of queue length. This survivor function can be used as an approximation to the CLR for a given finite queue size. In the next two subsections, we derive CLR of real-time voice and video traffic and stationary occupancy distribution of non-real-time queue length at the frame boundaries. We use the matrix-analytic technique to evaluate the stationary occupancy probabilities of the non-real-time queue.

5.2.1 CLR of Real-time traffic

Since real-time traffic is handled on a priority basis, its performance is not affected by the presence of non-real-time traffic. The real-time traffic as seen by the scheduler is the superposition of 2-state MMPP for voice traffic and 2-state MMPP for video traffic. Accordingly, the net arrival process of packets from real-time traffic can be represented by a 4-state MMPP[85]. If the infinitesimal generator matrix of the underlying Markov chain of the 2-state phase process for voice traffic is represented by Q_{vo} and for video traffic by Q_{vi}, then the infinitesimal generator matrix of the underlying phase process of superposed real-time traffic is given by,

$$Q_r \;=\; Q_{vi} \oplus Q_{vo} \tag{5.1}$$

where

$$Q_{vi} = \begin{pmatrix} -\sigma_{1,vi} & \sigma_{1,vi} \\ \sigma_{2,vi} & -\sigma_{2,vi} \end{pmatrix}$$

$$Q_{vo} = \begin{pmatrix} -\sigma_{1,vo} & \sigma_{1,vo} \\ \sigma_{2,vo} & -\sigma_{2,vo} \end{pmatrix}$$

and $\oplus$ stands for the Kronecker sum. The Kronecker sum for matrices A and B, is defined as

$$A \oplus B = (A \otimes I_B) + (I_A \otimes B)$$

where $\otimes$ represents the Kronecker product, which is defined as

$$C \otimes D = \begin{pmatrix} c_{11}D & c_{12}D & \cdots & c_{1m}D \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots & . \\ c_{n1}D & c_{n2}D & \cdots & c_{nm}D \end{pmatrix}$$

and I_A and I_B are the identity matrices of the same order as the matrices A and B respectively.

The corresponding arrival rate matrix of the superposed process can be represented by the following Kronecker sum.

$$\Lambda_r = \Lambda_{vi} \oplus \Lambda_{vo} \tag{5.2}$$

where,

$$\Lambda_{vi} = \begin{pmatrix} \lambda_{1,vi} & 0 \\ 0 & \lambda_{2,vi} \end{pmatrix}$$

$$\Lambda_{vo} = \begin{pmatrix} \lambda_{1,vo} & 0 \\ 0 & \lambda_{2,vo} \end{pmatrix}$$

Let π_r be the row vector of dimension 4 whose elements stand for the steady state probability of the phase process of the real-time traffic, i.e., the probability of being in phase i at any time. Thus,

$$\pi_r \cdot Q_r = 0 \tag{5.3}$$

$$\pi_r \cdot \underline{e} = 1$$

where $\underline{e}$ is the unit column vector of dimension 4.

Consider the epoch, just after serving the packets from both the real-time buffer and the non-real-time data queue. The number of packets in the real-time buffer will always be zero at these instants. Therefore, the steady state probabilities of cell loss (or cell loss rates, CLR's) of real-time packets can be evaluated in the following manner. Let τ be the frame time, $N_r(n\tau)$ be the number of real-time cell arrivals during n-th frame, and $J_r(n\tau)$ be the phase of the 4-state MMPP of real-time traffic(see Q_r) at the corresponding instants. We define

$$P_{r_{ij}}(k, n\tau) \quad = \quad P(N_r(n\tau) = k, J_r(n\tau) = j / J_r((n-1)\tau) = i) \tag{5.4}$$

Note that $P_{r_{ij}}(k, n\tau)$ is not dependent on the index n and can be evaluated, by using the recursion suggested in[86], based on an uniformization technique (during a frame period, the transition rate of the phase process of MMPP is in accordance with Q_r given in Equation 5.1). Let us denote the matrix with elements $P_{r_{ij}}(k, n\tau)$ by $P_r(k)$, i.e.,

$$P_r(k) \quad = \quad \left[P_{r_{ij}}(k, n\tau)\right] \tag{5.5}$$

which is a 4×4 matrix for each k. The real-time demand $f_r(k)$ is the marginal probability that $N_r(n\tau) = k$, which is found by averaging over the phase.

$$f_r(k) \quad = \quad \pi_r \cdot P_r(k) \cdot \underline{e} \tag{5.6}$$

Thus the cell loss probability for real-time traffic (CLP_{RT}), is given by

$$CLP_{RT} \quad = \quad \sum_{i=C+1}^{\infty} f_r(i) \tag{5.7}$$

Note that, if the real-time packets were queued in a finite buffer, then CLR_{RT} can be calculated only after modeling the queuing process of real-time traffic.

5.2.2 Queue Length of Non-real-time traffic

As explained in the previous section, $J_r(n\tau)$ represents the phase of the 4-state MMPP of the real-time traffic at the beginning of the n-th frame. Similarly, let $J_d(n\tau)$ stand for the phase of the 2-state MMPP representing the non-real-time traffic. The random variable representing the queue size of the non-real-time traffic($q((n+1)\tau)$) at time $(n+1)\tau$ depends on the queue size of the non-real-time traffic($q(n\tau)$) at time $n\tau$ and the phases of the 4-state MMPP of the real-time traffic and the 2-state MMPP of the non-real-time traffic. Thus we have an embedded Markov chain for the queue size of the non-real-time traffic, at the beginning of each frame time. Let $N_d(n\tau)$ be the random variable representing the number of non-real-time cell arrivals during n-th frame. Therefore, the equation governing the non-real-time queue process at the beginning of frame times after scheduling real-time packets and non-real-time packets can be given by,

$$q((n+1)\tau) \quad = \quad \left[q(n\tau) + N_d(n\tau) - [C - N_r(n\tau)]^+\right]^+ \qquad (5.8)$$

The evolution of the process described by Equation 5.8, is governed by a three dimensional embedded Markov chain whose components are the contents of the buffer, the phase of the 4-state MMPP of the real-time traffic and the phase of the 2-state MMPP of the non-real-time traffic. The transition probability of this chain is given by,

$$P_{d_{ijl,uvw}} =$$

$$P_d\left[q((n+1)\tau) = u, J_d((n+1)\tau) = v, J_r((n+1)\tau) = w / q(n\tau) = i, J_d(n\tau) = j, J_r(n\tau) = l\right] \qquad (5.9)$$

We now define the matrix

$$P_d = \left[P_{d_{ijl,uvw}}\right]$$

The remainder of this section is taken up with putting P_d into the canonical form

$$P_d = \begin{pmatrix} B_0 & B_1 & B_2 & B_3 & B_4 & \dots\dots \\ A_0 & A_1 & A_2 & A_3 & A_4 & \dots\dots \\ 0 & A_0 & A_1 & A_2 & A_3 & \dots\dots \\ 0 & 0 & A_0 & A_1 & A_2 & \dots\dots \\ . & . & . & . & . & \dots\dots \\ . & . & . & . & . & \dots\dots \end{pmatrix}$$

This form is readily solved using well-known matrix analytic techniques. We begin by defining the 8 × 8 sub-matrix

$$G_0 = \begin{pmatrix} P_{d_{011,011}} & P_{d_{011,012}} & P_{d_{011,021}} & P_{d_{011,022}} & \cdots & P_{d_{011,042}} \\ P_{d_{012,011}} & P_{d_{012,012}} & P_{d_{012,021}} & P_{d_{012,022}} & \cdots & P_{d_{012,042}} \\ P_{d_{021,011}} & P_{d_{021,012}} & P_{d_{021,021}} & P_{d_{021,022}} & \cdots & P_{d_{021,042}} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ P_{d_{042,011}} & P_{d_{042,012}} & P_{d_{042,021}} & P_{d_{042,022}} & \cdots & P_{d_{042,042}} \end{pmatrix}$$

The elements of this matrix are the probability of going from an empty buffer to an empty buffer with phase transitions. The elements of G_0 are given by the following expression.

$$P_{d_{0jl,0vw}} = \sum_{m=0}^{C}\sum_{i=0}^{C-m} P_{d_{jv}}(i)P_{r_{lw}}(m) + \sum_{m=C+1}^{\infty} P_{d_{jv}}(0)P_{r_{lw}}(m) \qquad (5.10)$$

where $P_{d_{jv}}(i)$ is the probability of number of non-real-time cell arrivals at the end of the frame period being equal to i and the phase of 2-state MMPP of non-real-time traffic at the end of the frame period is v given that the phase at the beginning of the frame period is j. This probability can be obtained using the same recursive method[86] as employed in the case of real-time traffic. Equation 5.10. can be written in a matrix form using the Kronecker product as shown below.

$$G_0 = \sum_{m=0}^{C}\sum_{i=0}^{C-m} P_d(i) \otimes P_r(m) + \sum_{m=C+1}^{\infty} P_d(0) \otimes P_r(m) \qquad (5.11)$$

where $P_r(m) = [P_{r_{lw}}(m)]$ and $P_d(i) = \left[P_{d_{jv}}(i)\right]$.

In the same manner, we define the sub-matrices which involve a transition from an empty buffer to an non-empty buffer.

$$G_i = \left[P_{d_{0jl,ivw}}\right] \;\; for \;\; i \geq 1: \quad 1 \leq j,v \leq 2 \;\; : \;\; 1 \leq l,w \leq 4$$

with

$$P_{d_{0jl,ivw}} = \sum_{m=0}^{C} P_{d_{jv}}(C-m+i)P_{r_{lw}}(m) + \sum_{m=C+1}^{\infty} P_{d_{jv}}(i)P_{r_{lw}}(m)$$

In matrix form,

$$G_i = \sum_{m=0}^{C} P_d(C-m+i) \otimes P_r(m) + \sum_{m=C+1}^{\infty} P_d(i) \otimes P_r(m)$$

The elements of G_i are obtained by looking into the mutually exclusive cases of leaving behind i fixed size packets in the queue. after scheduling the transmission of real-time packets and non-real-time packets from the non-real-time queue at the beginning of the current frame. given that there were no non-real-time packets in the non-real-time queue at the beginning of the previous frame. Thus. this defines the first row of the transition probability matrix $P_{d_{ijl,uvw}}$. In a similar manner. we write down the elements of other sub-matrices for the transition from an non-empty buffer to non-real-time queues of other states.

$$D = \left[P_{d_{1jl.1vw}}\right] \qquad 1 \leq j.v \leq 2 \ \ : \ \ 1 \leq l,w \leq 4$$

where

$$P_{d_{1jl.1vw}} = \sum_{m=0}^{C} P_{d_{jv}}(C-m)P_{r_{lw}}(m) + \sum_{m=C+1}^{\infty} P_{d_{jv}}(0)P_{r_{lw}}(m)$$

In matrix form.

$$D = \sum_{m=0}^{C} P_d(C-m) \otimes P_r(m) + \sum_{m=C+1}^{\infty} P_d(0) \otimes P_r(m)$$

$$\left[P_{d_{1jl,ivw}}\right] = G_{i-1} \qquad i \geq 2: \quad 1 \leq j,v \leq 2 \ \ : \ \ 1 \leq l,w \leq 4$$

$$U_i = \left[\sum_{m=0}^{C-i} \sum_{n=0}^{C-m-i} P_{d_{jv}}(n)P_{r_{lw}}(m)\right] \qquad for \quad 1 \leq i \leq C$$

In matrix form.

$$U_i = \sum_{m=0}^{C-i} \sum_{n=0}^{C-m-i} P_d(n) \otimes P_r(m) \qquad for \quad 1 \leq i \leq C$$

$$E_i = \left[\sum_{m=0}^{C-i} P_{d_{jv}}(C-m-i)P_{r_{lw}}(m)\right] \qquad for\ \ 1 \le i \le C-1$$

In matrix form,

$$E_i = \sum_{m=0}^{C-i} P_d(C-m-i) \otimes P_r(m) \qquad for\ \ 1 \le i \le C-1$$

Thus, we have the following form for the transition probability matrix P_d,

$$P_d = \begin{pmatrix}
G_0 & G_1 & G_2 & G_3 & G_4 & \cdots & G_{C-1} & G_C & G_{C+1} & \cdot & G_{2C-1} & G_{2C} & \cdots \\
U_1 & D & G_1 & G_2 & G_3 & \cdots & G_{C-2} & G_{C-1} & G_C & \cdot & G_{2C-2} & G_{2C-1} & \cdots \\
U_2 & E_1 & D & G_1 & G_2 & \cdots & G_{C-3} & G_{C-2} & G_{C-1} & \cdot & G_{2C-3} & G_{2C-2} & \cdots \\
U_3 & E_2 & E_1 & D & G_1 & \cdots & G_{C-4} & G_{C-3} & G_{C-2} & \cdot & G_{2C-4} & G_{2C-3} & \cdots \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\
U_{C-1} & E_{C-2} & E_{C-3} & E_{C-4} & E_{C-5} & \cdots & D & G_1 & G_2 & \cdot & G_C & G_{C+1} & \cdots \\
U_C & E_{C-1} & E_{C-2} & E_{C-3} & E_{C-4} & \cdots & E_1 & D & G_1 & \cdot & G_{C-1} & G_C & \cdots \\
\mathbf{0} & U_C & E_{C-1} & E_{C-2} & E_{C-4} & \cdots & E_2 & E_1 & D & \cdot & G_{C-2} & G_{C-1} & \cdots \\
\mathbf{0} & \mathbf{0} & U_C & E_{C-1} & E_{C-3} & \cdots & E_3 & E_2 & E_1 & \cdot & G_{C-3} & G_{C-2} & \cdots \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdots
\end{pmatrix}$$

We further group these sub-matrices in the following manner.

$$B_0 = \begin{pmatrix}
G_0 & G_1 & G_2 & \cdot\cdot & G_{C-1} \\
U_1 & D & G_1 & \cdot\cdot & G_{C-2} \\
U_2 & E_1 & D & \cdot\cdot & G_{C-3} \\
U_3 & E_2 & E_1 & \cdot\cdot & G_{C-4} \\
\cdot & \cdot & \cdot & \cdot\cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot\cdot & \cdot \\
U_{C-1} & E_{C-2} & E_{C-3} & \cdot\cdot & D
\end{pmatrix}
B_i = \begin{pmatrix}
G_{iC} & G_{iC+1} & \cdot\cdot & G_{(i+1)C-1} \\
G_{iC-1} & G_{iC} & \cdot\cdot & G_{(i+1)C-2} \\
G_{iC-2} & G_{iC-1} & \cdot\cdot & G_{(i+1)C-3} \\
G_{iC-3} & G_{iC-2} & \cdot\cdot & G_{(i+1)C-4} \\
\cdot & \cdot & \cdot\cdot & \cdot \\
\cdot & \cdot & \cdot\cdot & \cdot \\
G_{(i-1)C+1} & G_{(i-1)C+2} & \cdot\cdot & G_{iC}
\end{pmatrix}$$

$for\ \ i \ge 1$.

$$A_0 = \begin{pmatrix}
U_C & E_{C-1} & E_{C-2} & \cdot\cdot & E_1 \\
\mathbf{0} & U_C & E_{C-1} & \cdot\cdot & E_2 \\
\mathbf{0} & \mathbf{0} & U_C & \cdot\cdot & E_3 \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdot\cdot & E_4 \\
\cdot & \cdot & \cdot & \cdot\cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot\cdot & \cdot \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \cdot\cdot & U_C
\end{pmatrix}
A_1 = \begin{pmatrix}
D & G_1 & \cdot\cdot & G_{C-1} \\
E_1 & D & \cdot\cdot & G_{C-2} \\
E_2 & E_1 & \cdot\cdot & G_{C-3} \\
E_3 & E_2 & \cdot\cdot & G_{C-4} \\
\cdot & \cdot & \cdot\cdot & \cdot \\
\cdot & \cdot & \cdot\cdot & \cdot \\
E_{C-1} & E_{C-2} & \cdot\cdot & D
\end{pmatrix}$$

and $A_{i+1} = B_i$ for $i \geq 1$. Note that the matrices A_i and B_i are of the order of 8·C. Let Π_{uvw} be the joint steady state probability of queue size of non-real-time traffic being equal to u, the phase of 4-state MMPP of real-time traffic being in state v and the phase of 2-state MMPP of non-real-time traffic being in state w. The steady state joint probability vector of number of non-real-time packets in the queue, the phase of 4-state MMPP of real-time traffic and the phase of 2-state MMPP of non-real-time traffic, being at the embedded points (i.e., at the beginning of frame times) can be written as,

$$\Pi_d \quad = \quad ([\Pi_{0vw}], [\Pi_{1vw}], [\Pi_{2vw}], ...) \tag{5.12}$$

where each sub-vector $[\Pi_{uvw}]$ is a vector of dimension 8, with the elements Π_{uvw} having the queue size of non-real-time traffic being u, but with different phases of 4-state MMPP of real-time traffic and 2-state MMPP of non-real-time traffic, arranged in a lexicographical order. We use the matrix-geometric technique to solve for the steady state probabilities Π_{uvw} [37][79][81][82][86][95]. Since, $A_{i+1} = B_i$ for $i \geq 1$, this relationship helps to simplify the calculation of Π_{uvw} as shown in the following section.

5.3 Steps For the Computation of Stationary Probability Vector

Each row in the composed matrix, i.e., P_d (specified in page 56), corresponds to a level as in the case of M/G/1 type Markov chains, i.e., level $\underline{l}$ is given by,

$$\underline{l} = \{(l * C + k, j, v); 0 \leq k \leq C - 1, 1 \leq j \leq m, 1 \leq v \leq n\} \qquad for \qquad l \geq 0$$

where m and n stand for the number of phases of rt MMPP and nrt data MMPP, respectively.

The Markov chain has skip-free to the left property. Therefore, we can proceed with the analysis as in the case of M/G/1 type Markov chains by considering the first passage times between various pairs of levels.

Now, the steady state vector $\underline{\mathbf{x}_l}$ is a row vector corresponding to each level. Within this vector each element corresponds to the steady state probability of the system with which we started, i.e.. Π_{ijv}.

We briefly outline the solution technique here.

Step 1: We have to solve for G from the following non-linear matrix equation.

$$G = \sum_{v=0}^{\infty} A_v G^v$$

Let $\underline{g}$ be the invariant probability vector of the matrix G.

Step 2: We determine the vector

$$\underline{\beta^*} = \sum_{v=1}^{\infty} v A_v \underline{e}$$

where, $\underline{e}$ is the column vector of 1s with dimension of $C \cdot m \cdot n$.

Step 3: Defining $A = \sum_{v=0}^{\infty} A_v$ and $\tilde{G}$ and $\Delta(\underline{\beta^*})$ as diagonal matrix of order $Cmn \times Cmn$ with the components of $\underline{\beta^*}$ as its diagonal elements. We calculate.

$$\underline{\phi} = (I - G + \tilde{G}) * (I - A + \tilde{G} - \Delta(\underline{\beta^*})\tilde{G})^{-1}\underline{e}$$

Step 4:

Let $\underline{\kappa}$ be invariant probability vector of $\sum_{v=0}^{\infty} B_v G^v$. Since $A_{v+1} = B_v$ for $v \geq 1$. we can write $\sum_{v=0}^{\infty} B_v G^v = I - (A_0 G^{-1}) - A_1 + B_0$.

Let $\underline{\kappa^*} = \underline{e} + \sum_{v=1}^{\infty} B_v \sum_{r=0}^{v-1} G^r \underline{\phi}$. We calculate.

$$\underline{x_0} = (\underline{\kappa \kappa^*})^{-1} \underline{\kappa}$$

Step 5:

Finally, by using the recursion suggested by V.Ramaswami, we can obtain $\underline{x_i}$'s from the following equations,

$$\underline{x_i} = \underline{x_0} A_{i+1} + \sum_{v=1}^{i-1} \underline{x_v} A_{i+1-v} \quad for \quad i \geq 1$$

5.4 Case Study: I

The CLR of rt-VBR voice and video traffic and stationary distribution of nrt-VBR queue length at the frame boundaries are derived using the analysis presented in the previous section. In the following, we present some examples with illustrative results for a typical TDMA frame duration of 1 millisecond. We take into account the overhead associated with the header information and assume the fixed size packet(WATM cell) of 54 bytes to carry an ATM SDU of 48 bytes. The WATM cell contains a header of 6 bytes. An additional byte as compared to the standard 53-byte ATM cell is for the data link error control mechanism[26]. As illustrative examples[92], two different uplink capacities are considered, $C_u = 5Mbps$ and $C_u = 12.5Mbps$. With a 1 millisecond frame length these rates yield $C = 11$ and $C = 28$ slots per frame, respectively. Our numerical results involve various mixes of the traffic classes. The salient characteristics of voice, video and data traffic are shown in Table 5.1. We consider the delay requirements of both 64 Kbps voice and 384 kbps video sources to be nearly the same, as listed in [23], page no. 39. That is why we consider both voice and video at the same priority level, without discriminating between voice and video. Two different mixes of the traffic types, "data dominant" and "balanced" respectively are shown in Table 5.2. The results of calculations are summarised in Tables 5.3 and 5.4 for $C_u = 5$ and 12.5 Mbps, respectively. The two performance criteria are Statistical Multiplexing Gain(SMG) and Cell Loss Ratio(CLR), respectively. The SMG is defined as $SMG = ((N_{voi} * 64Kbps) + (N_{vid} * 384Kbps) + (N_{data} * 128Kbps))/(C_u)$. Thus, this expression compares the gain that we can have against fixed rate PCR(Peak cell rate) CBR service(or equivalently circuit switching). The Peak-to-Average ratio(PAR) of the on-off/mini-on-off sources is defined as the ratio of peak bit rate to average bit rate.

Since our objective is to satisfy the QoS requirements of CLR and maximum end-to-end delay requirement of rt-connections and CLR of nrt-connections, we choose

Traffic Type	Model Details	Peak bit rate for On-Off/ mini-On-Off Source	PAR
RT Voice	Single On-Off Source	64 Kbps	2.5
RT Video	6 mini On-Off Sources	64 Kbps	5.0
NRT Data	Single On-Off Source	128 Kbps	200

Table 5.1: Statistical characteristics of different source types.

Traffic Type	Data Dominant	Balanced
RT Voice	20%	20%
RT Video	10%	30%
NRT Data	70%	50%

Table 5.2: Different cases of traffic mix.

Load	Case of Traffic Mix	N_{voi}	N_{vid}	N_{data}	SMG	CLR of RT connections
90%	Data Dominant	29	4	4158	107.12	6.654×10^{-5}
80%	Data Dominant	26	4	3696	95.26	3.58×10^{-5}
50%	Data Dominant	16	2	2310	59.49	2.941×10^{-7}
50%	Balanced	16	8	1650	43.06	5.616×10^{-5}

Table 5.3: CLR and Statistical Multiplexing Gain(SMG) of different cases of traffic mix for the capacity of $C_u = 5Mbps$ with 1 msec WATM uplink frame duration.

CLR of rt-VBR connections to be less than 10^{-4} at the cluster. As previously discussed, our scheduling policy gives priority to rt packets. Furthermore, those rt cell arrivals in a frame period exceeding the number of allocated slots in a frame are lost. The delay of rt packets is bounded by a frame duration at the cluster. In Table 5.3 and Table 5.4, we show the CLR of rt packets at the cluster. We observed for the cases of load 95% through 35%, the voice and video dominant cases(80% rt traffic and 20% nrt traffic) lead to a CLR higher than 10^{-4} for the capacity of $C_u = 5Mbps$. Similarly, for the cases of load 95% through 60%, the voice and video dominant cases lead to a CLR of much higher than 10^{-4} for the capacity of $C_u = 12.5Mbps$ (not listed in the Tables).

In the following Figures 5.2. and 5.3 we show the logarithm of survivor function of nrt data queue occupancy vs. nrt data queue size corresponding to Table 5.3. From these figures, we can determine the nrt data queue size at the user terminals in order to ensure a particular value of CLR. For example to ensure a CLR of 10^{-8} for nrt data traffic with the assumed traffic model, we have to have a nrt data queue size of 6746, with the radio capacity of 5Mbps devoted for the uplink under 90% load of data dominant case shown in Figure 5.3. Similarly, in Figures 5.4. and 5.5, we show the logarithm of survivor function of nrt data queue occupancy vs. nrt data queue size corresponding to Table 5.4. We can choose the nrt data queue size in order to

Load	Case of Traffic Mix	N_{voi}	N_{vid}	N_{data}	SMG	CLR of RT connections
90%	Data Dominant	75	12	10584	109.13	7.613×10^{-9}
90%	Balanced	75	37	7560	78.935	8.1685×10^{-5}
80%	Data Dominant	67	11	9408	97.02	8.14×10^{-10}
80%	Balanced	67	33	6720	70.17	1.264×10^{-5}
50%	Data Dominant	42	7	5880	60.64	3.746×10^{-14}

Table 5.4: CLR and Statistical Multiplexing Gain(SMG) of different cases of traffic mix for the capacity of $C_u = 12.5Mbps$ with 1 msec WATM uplink frame duration.

meet a particular CLR of nrt data connections.

From Figure 5.4. we can see that at 80% load. the curve of logarithm of survivor function of non-real-time data queue length for data dominant case is above the curve of logarithm of survivor function of non-real-time data queue length for balanced case. However. in Figure 5.5. it is the other way around. This leads to the following intriguing argument. At higher loads. for the same load. as the proportion of real-time traffic(voice and video) increases. the non-real-time data queue length tends to increase. This can be attributed to the priority nature of the system. This behaviour can be studied exactly. by developing sensitivity analysis of logarithm of the survivor function of non-real-time data queue length.

In Figure 5.6. we show the logarithm of survivor function of *aggregate* nrt data queue as seen by the scheduler and the *individual* terminal nrt data queue for the data dominant case of load 90%[95]. The simulation for this case is performed with 6 WATM terminals. with each having 12 voice sources. 2 video sources and 1764 data sources. For our simulations. we choose a confidence level of 95%; all of the simulation points are within the confidence interval of 5% with this confidence level. For the details of calculating the confidence interval for occupancy probability. refer Section 4.1. Each voice. video or data source is represented by an ON-OFF model

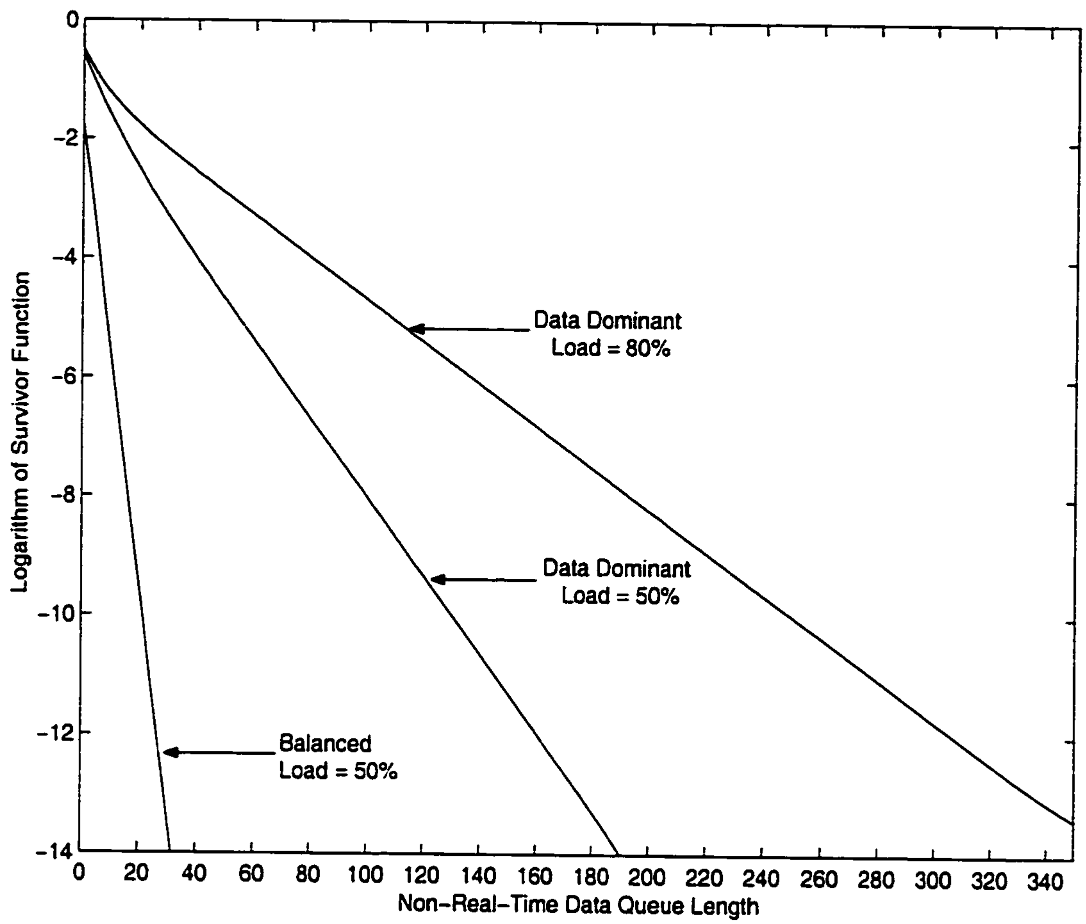

Figure 5.2: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for $C_u = 5Mbps$ and for a load 80% and 50% of the cases corresponding to Table 5.3 (Analytical results).

or a superposition of ON-OFF mini-sources as shown in Table 5.1. Curves #1 and #3 show the simulation results for the *aggregate* queue as seen by the scheduler and the *individual* queue at the terminal. respectively. when traffic is represented by its original ON-OFF sources or mini-sources.

The 12 voice sources can also be approximated by a 2-state MMPP. in the same manner as was done for all previous cases. Similarly. the 2 video sources are approximated by one 2-state MMPP and the 1764 data sources approximated by another 2-state MMPP. Curve #2 shows the analytical results for the *aggregate* queue

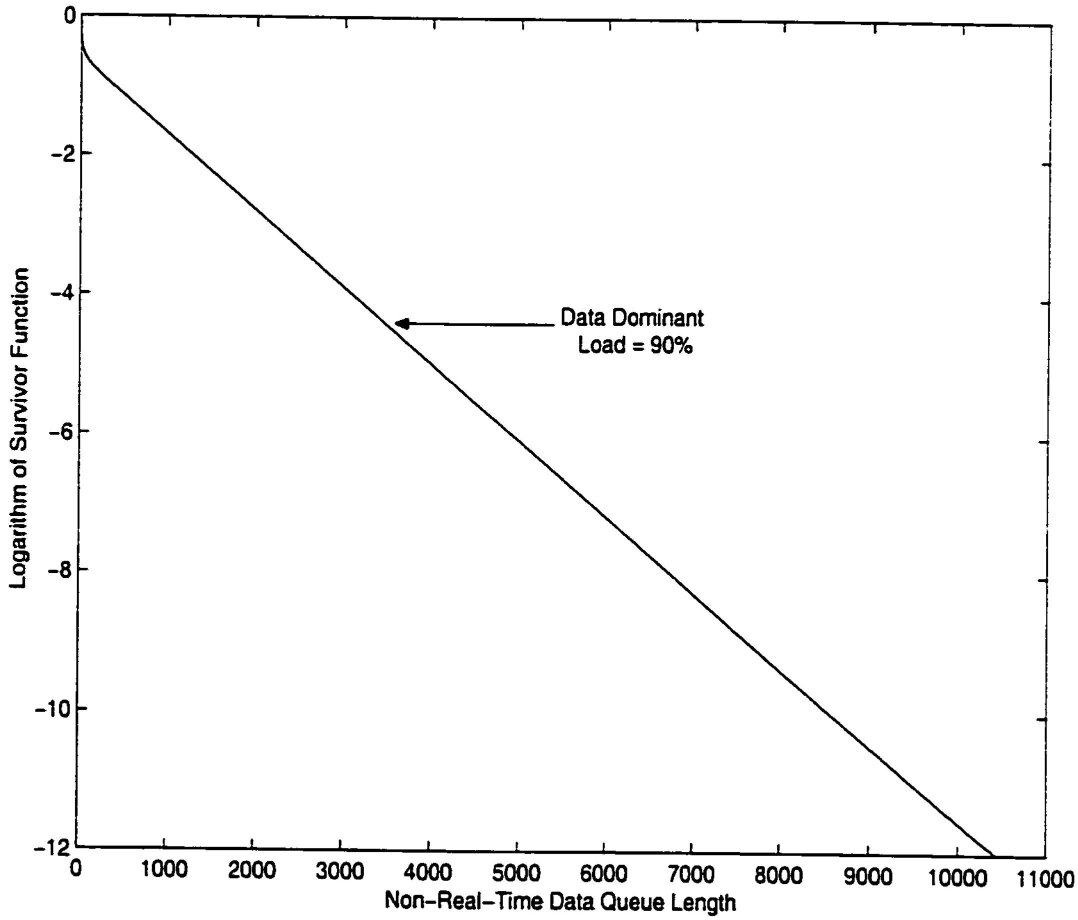

Figure 5.3: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for $C_u = 5Mbps$ and for a load 90% of data dominant case corresponding to Table 5.3 (Analytical results).

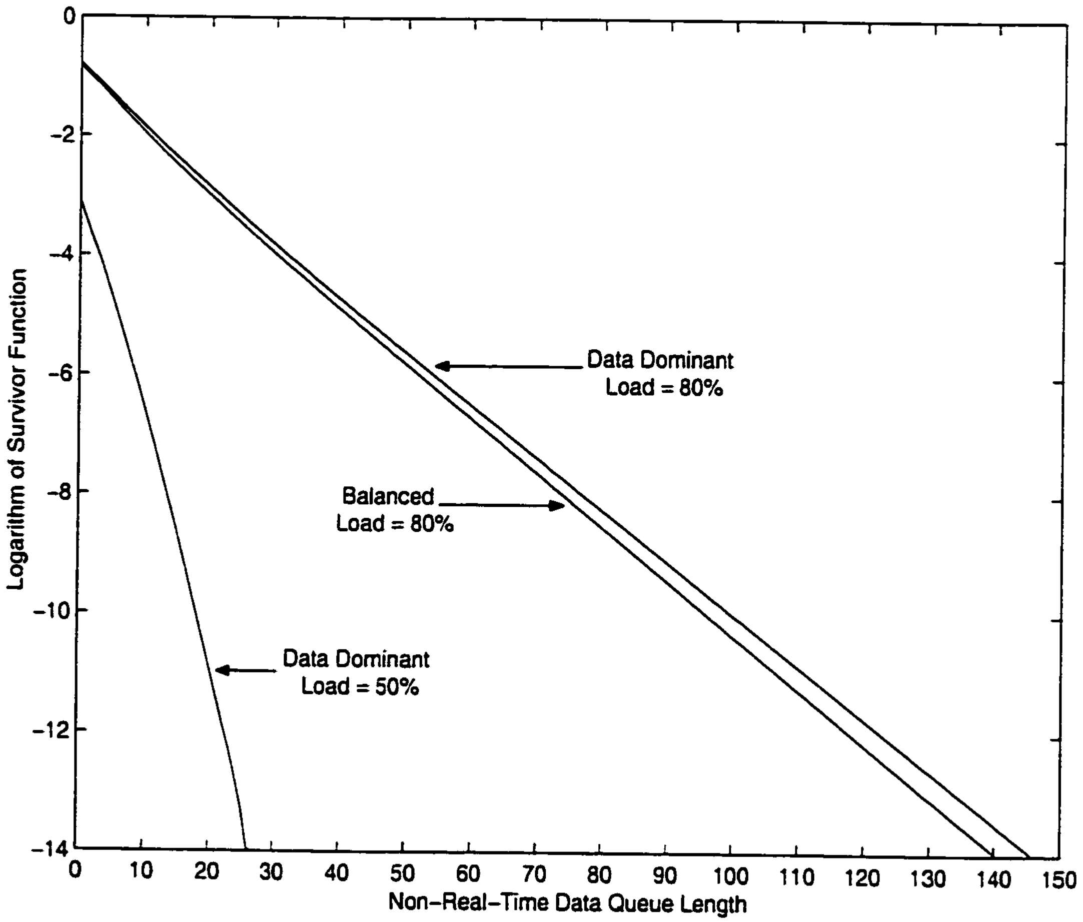

Figure 5.4: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for $C_u = 12.5Mbps$ and for a load 80% and 50% of the cases corresponding to Table 5.4 (Analytical results).

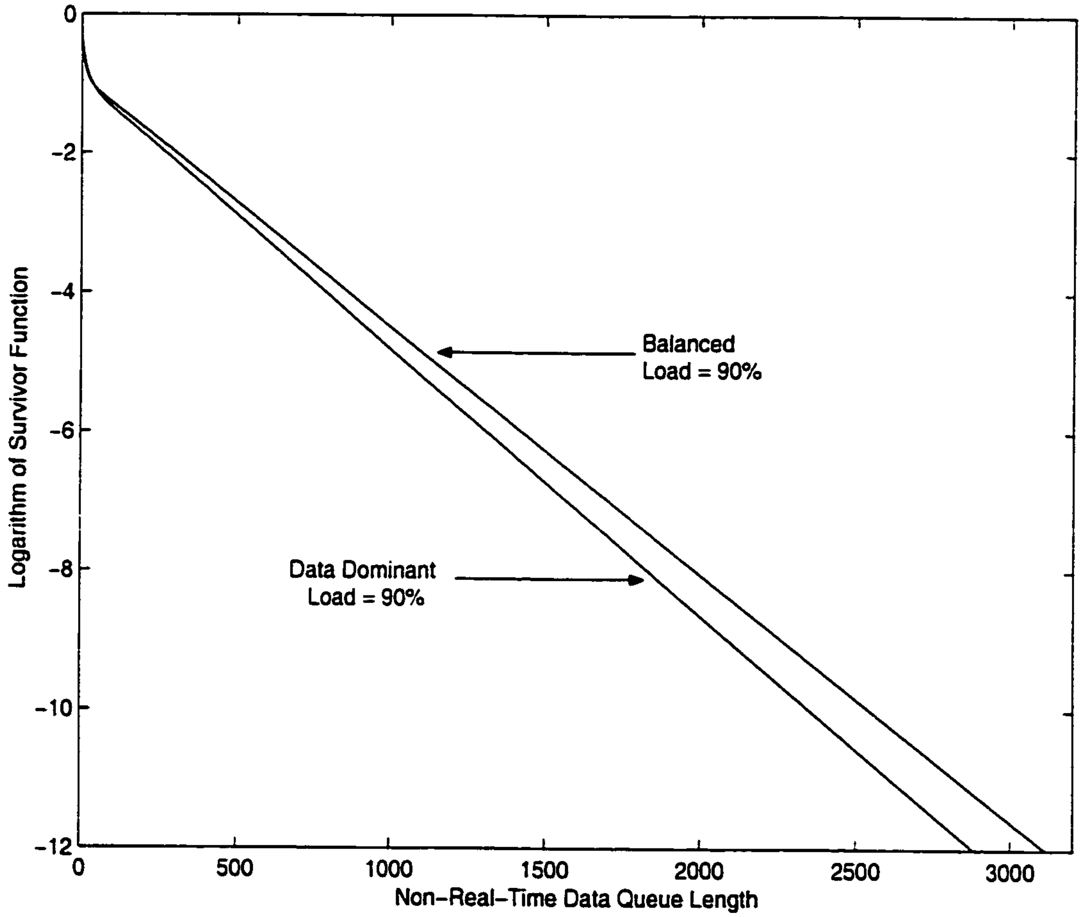

Figure 5.5: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for $C_u = 12.5Mbps$ and for a load 90% of data dominant case corresponding to Table 5.4 (Analytical results).

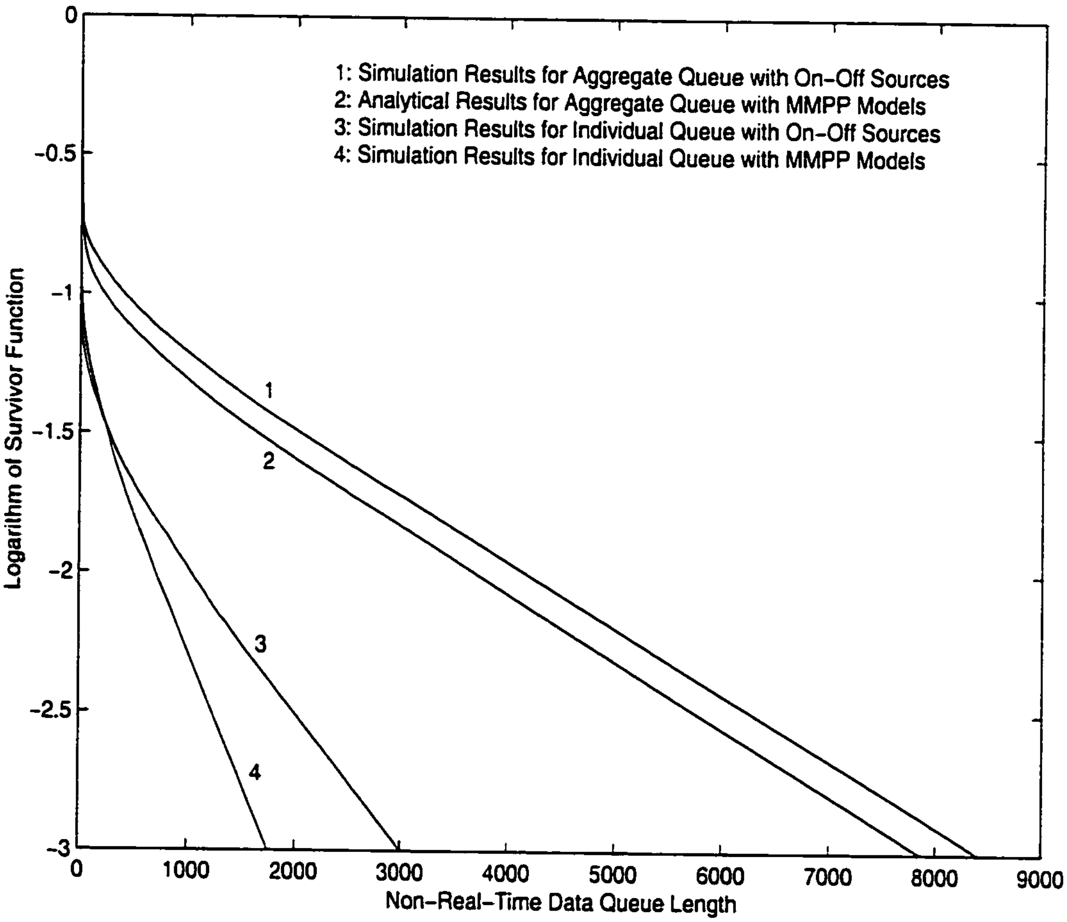

Figure 5.6: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for $C_u = 12.5Mbps$ and for a load 90% of data dominant case.

as seen by the scheduler when traffic is represented by three MMPP models for *aggregate* voice, video and data. Curve #4 shows the simulation results for the *individual* queue at the terminal with three MMPP models.

From Curves #1 and #2, we can see that the analytical results based on MMPP model are in close agreement to the simulation results based on ON-OFF models for the behavior of the *aggregate* queue[53]. A similar observation is applied to the behavior of the *individual* queue at the terminal by examining Curves #3 and #4. In other words. the approximation of an *aggregate* of homogeneous ON-OFF sources by a 2-state MMPP is well justified by simulation results. and the analytical model can be used as a fast way to dimension the queues.

Curves #2 and #4 (or Curves #1 and #3) also indicate a good approximation to estimate the queue size of each terminal from the size of the *aggregate* queue. For example. for a CLR of 1E-3. Curve # 2 (analytical results) shows that the *aggregate* queue of 6 terminals needs 7849 packets. Based on this result. the queue size of each terminal would be $\lceil 7849/6 \rceil = 1309$ packets while Curve 4 indicates the *individual* queue size of 1749 packets. Thus. we can use the aggregated queue size corresponding to a particular CLR value for individual terminals. which would provide much better CLR. or we can use aggregated queue size divided by the number of terminals. as a conservative approach (of course CLR performance will be worse).

5.5 Case Study: II

As another illustrative example[93], we consider an uplink frame of 24ms with a capacity of 64 time-slots per frame. which is typical of the satellite MF-TDMA system. Each time-slot can accomodate a fixed-size cell of 53 bytes(i.e.. ATM cell). Table 5.1 summarizes the statistical characteristics of voice. video and data traffic. In order to examine the effects of traffic mix on the system performance. we consider 3 different scenarios shown in Table 5.5 . The results on the CLR of real-time traffic for different

Traffic Type	Data Dominant	Video Dominant	Voice Dominant
RT Voice	20%	10%	70%
RT Video	10%	70%	10%
NRT Data	70%	20%	20%

Table 5.5: Different cases of traffic mix.

Load	Case of Traffic Mix	N_{voi}	N_{vid}	N_{data}	S.M.G.	CLR of RT traf.
78.5%	Data Dom.	6	1	895	102.0	$1.6976E-13$
	Voice Dom.	22	1	256	30.57	$7.44E-03$
	Video Dom.	4	7	256	31.58	$1.186E-02$
52.5%	Data Dom.	4	1	560	63.96	$1.7944E-17$
	Voice Dom.	14	1	160	19.25	$5.1755E-06$
	Video Dom.	2	5	160	19.92	$1.84E-05$

Table 5.6: CLR and Statistical Multiplexing Gain(S.M.G.) of different cases of traffic mix for the uplink capacity of C = 64 slots per 24 ms.

cases of traffic mix and load are given in Table 5.6. The statistical multiplexing gain(S.M.G.) is calculated as $S.M.G. = ((N_{voi} * 64Kbps) + (N_{vid} * 384Kbps) + (N_{data} * 128Kbps))/(C * 53 * 8/24ms)$. The performance of non-real-time traffic (e.g., data) for the cases shown in Table 5.6 is evaluated using the analytical tool.

In Figures 5.7, and 5.8, the logarithm of survivor function is plotted against the data queue length for different loads of 0.525, and 0.785 respectively. We find that even though the data load is higher in the data dominant case as compared to the video dominant case for the total load of 0.525, the queue occupancy is higher for the video dominant case due to high burstiness of real-time video traffic(Figure 5.7). On the other hand, at a higher load of 0.785, the data dominant case has a higher data queue occupancy than the video dominant case. The effect of burstiness due to video traffic on the data queue diminishes as the total load increases(Figure 5.8). We can approximate the tail probabilities for the CLR of non-real-time data traffic. Thus, for a particular CLR of non-real-time data, we can obtain the corresponding buffer size from these graphs. For example, to achieve a CLR of $6.34E - 07$ of non-real-time data traffic, the data buffer size has to be 4736 in the data dominant case, 3255 in the video dominant case and 2176 in the voice dominant case(Figure 5.8).

When the number of slots in a frame is increased to the order of 4096, the convergence of iterative solution(matrix geometric), is extremely slow. Although, for $C = 64$, the analytical results could be obtained within half an hour or so at higher loads(near 0.9), the convergence for the case of $C = 4096$, was taking of the order of days. For the purpose of comparison with simulation[91], we present the analytical and simulation results for the case of $C = 10$ and for the loads varying between 0.9 and 0.57, as shown in Figure 5.9. For our simulations, we choose a confidence level of 95%; all of the simulation points are within the confidence interval of 5% with this confidence level. For the details of calculating the confidence interval for occupancy probability, refer Section 4.1. The analytical results are in excellent agreement with simulation as shown in Figure 5.9.

In the next chapter, we derive the covariance function of arrival process of RT traffic and NRT traffic, to a downlink/downstream link. Using the covariance function values at different lags and mean arrival rate, we obtain the parameters for approximating the 2-state MMPP. Then we use the analytical model of uplink discussed in this chapter, to obtain the CLR of RT and NRT traffic at the downlink.

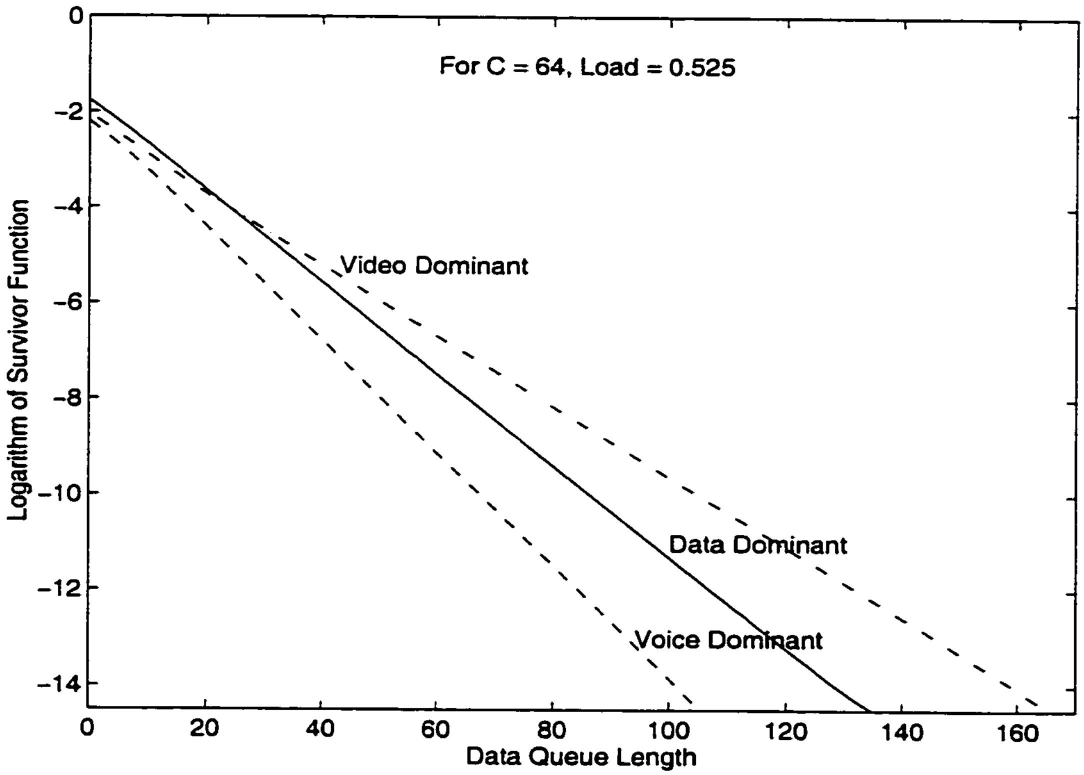

Figure 5.7: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for load = 0.525 (Analytical results).

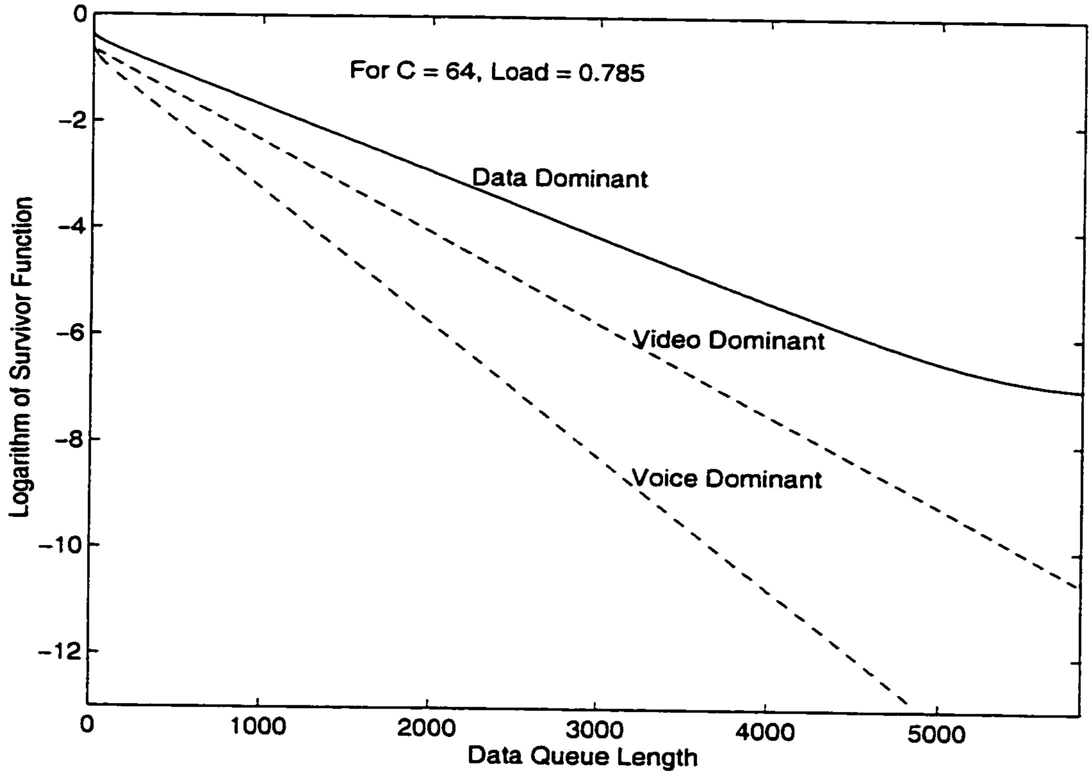

Figure 5.8: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for load = 0.785 (Analytical results).

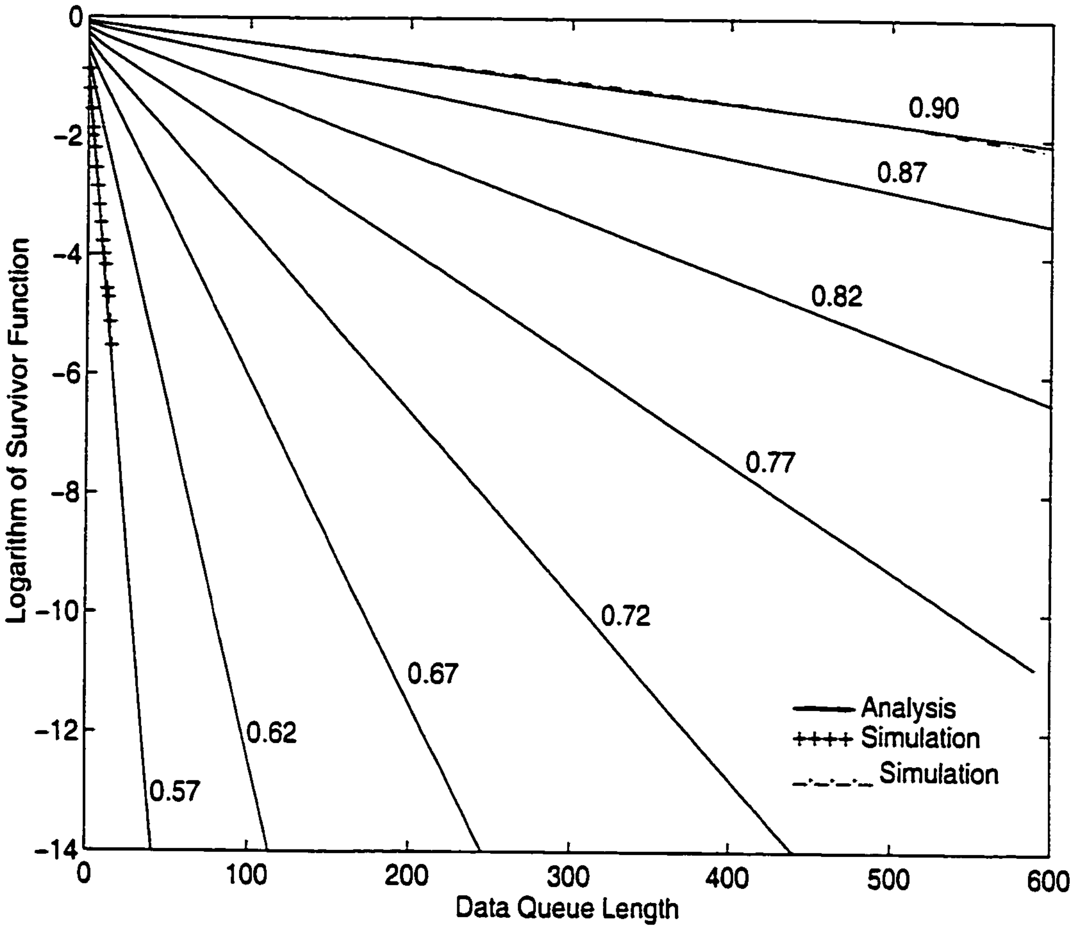

Figure 5.9: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length with the real-time traffic load being 0.5172 in all cases and with varying data traffic load ($\lambda_{1,vi} = \lambda_{1,vo} = 30$, $\sigma_{1,vi} = \sigma_{2,vi} = 2.4287$, $\lambda_{2,vi} = \lambda_{2,vo} = 220$, $\sigma_{1,vo} = \sigma_{2,vo} = 6.7115$, $\lambda_{1,data} = 20$ and $\lambda_{2,data}$ is varying) for the Uplink of capacity $C = 10$.

Chapter 6

Performance Evaluation of Priority Based Scheduling at the Downstream Link

In the simple case of a two hop connection, traffic from the user terminals may simply be switched from the local base station to a remote base station through the switch for WATM networks. In this case the output queue of the switch corresponds to the downlink of the remote base station. In all other cases, the output of the switch will be the next downstream link after the uplink. In the OBS based Satellite networks, the traffic arriving at the on-board switch will be switched to the appropriate downlink. Since rt-VBR has to be given priority over nrt-VBR, the same frame based scheduling is used over the downlinks/ downstream links. Our main objective is to characterise the arrival processes of real-time and non-real-time traffic at the downlink/ downstream link, so that performance at any downlink/ downstream link can be studied. The typical situation has been presented in Chapter 3, in Figure 3.5.

6.1 Analytical Modeling of Downstream Link

To characterise the arrival process of real-time and non-real-time traffic at the downlink/ downstream link, we only need to consider the aggregate buffer of real-time traffic which stores cells arriving in the previous frame and the aggregate queue of non-real-time traffic, from the user terminals to the base station or from the user terminals to the scheduler on-board. In the previous chapter, assuming voice and video traffic have the same delay and loss requirements, we considered a 4-state MMPP representation of aggregate voice and video traffic. For the non-real-time traffic we considered a 2-state MMPP. Let us consider the downstream queue of Figure 3.5 shown in Chapter 3. The fixed size packets departing from the input port choose the downstream output port according to a uniform distribution. This downstream output port would actually correspond to the buffer and/or queue at the base station for the downlink traffic, if the output port of ATM switch is directly connected to a base station or the output port of the ATM switch, with the scheduler operating on a TDMA frame basis. In the case of OBS based satellite network, the output port would correspond to the downlink. In that case, the queueing model (Figure 5.1) of the scheduler is applicable for the downlink/ downstream link queueing at the base station or the output port of the ATM switch if the arrival process to the downstream queue can be approximated by MMPPs. Thus, our uplink modeling analysis discussed in previous chapter on the steady state distribution of the above queueing structure can be used to analyse the queueing characteristics at the downstream queue.

An MMPP is characterised by the transition rate matrix of the underlying phase process and a diagonal arrival rate matrix. Let Q_r(of order m) and Q_d(of order n) be the transition rate matrices of the underlying phase process of real-time and non-real-time traffic respectively, as discussed in previous chapter. Similarly, let Λ_r(of order m) and Λ_d(of order n) be the diagonal arrival rate matrices of the real-time and non-real-time traffic respectively, again as discussed in previous chapter. Let τ be the frame time. Let $P(N_r'(t) = k, J_r(t) = j/J_r(0) = i)$ be the conditional probability of the number of real-time arrivals being equal to k by time t and the phase of real-time process being in state j at time t, given the phase of the counting process $N_r'(t)$ at time 0 being in state i. Similarly, let $P(N_d'(t) = k, J_d(t) = j/J_d(0) = i)$ be the conditional probability of the number of non-real-time arrivals being equal to k by time t and the phase of non-real-time MMPP being in state j at time t, given that the phase of the counting process $N'_d(t)$ at time 0 being in state i.

In the next section we present the analysis for evaluating the covariance of the number of real-time arrivals in frame i and number of real-time arrivals in frame $(i+j)$ at the downlink/ downstream link from one single uplink or input/base station(see Figure 3.5). In the subsequent section we present the analysis for evaluating the covariance of the number of non-real-time arrivals in frame i and number of non-real-time arrivals in frame $(i+j)$ at the downlink/ downstream link from one single uplink or input/base station.

6.2 Covariance Function of Real-time Arrival Process to the Downlink from an Uplink

The notable work on departure process analysis for the case of an ATM multiplexer with correlated inputs can be found in [70][72][73][74]. Our analysis is quite different from these. due to the nature of the frame structure and priority based scheduling[94][96]. Additionally, we focus our attention on the arrival process of traffic to the downstream link after being switched. Since real-time traffic is handled on a priority basis. its performance is not affected by the presence of non-real-time traffic.

Let $N_{a,RT}(i\tau)$ be the number of real-time arrivals from the uplink to the particular downlink of interest during i-th frame. Let $N_{d,RT}(i\tau)$ be the total number of real-time departures from the uplink during i-th frame. Since. the real-time packets exceeding the capacity during each frame are lost. we can write down the probability distribution of $N_{d,RT}(i\tau)$ as shown below.

$$P\{N_{d,RT}(i\tau) = k, J_r(i\tau) = l/J_r((i-1)\tau) = m\}$$

$$= \begin{cases} P\{N_r(i\tau) = k, J_r(i\tau) = l/J_r((i-1)\tau) = m\} & 0 \leq k \leq C-1 \\ \sum_{i=C}^{\infty} P\{N_r(i\tau) = i, J_r(i\tau) = l/J_r((i-1)\tau) = m\} & k = C \end{cases}$$

The probability expression on the right hand side of the equation represents the probability of number of arrivals during i-th frame due to the real-time MMPP, at the uplink. Let $N_r(i\tau)$ be the random variable representing number of real-time arrivals during i-th frame. Since, out of k departing real-time packets from the uplink, i of them can be destined to the particular downlink of interest, with binomial probability $\begin{pmatrix} k \\ i \end{pmatrix} \left(\frac{1}{N}\right)^i \left(1 - \frac{1}{N}\right)^{k-i}$. we can write

$$P\{N_{a,RT}(i\tau) = k, J_r(i\tau) = l/J_r((i-1)\tau) = m\}$$

$$= \begin{cases} \sum_{l=0}^{C} P\{N_{d,RT}(i\tau) = l, J_r(i\tau) = l/J_r((n-1)\tau) = m\} \left(1 - \frac{1}{N}\right)^l & k = 0 \\ \sum_{l=k}^{C} \begin{pmatrix} l \\ k \end{pmatrix} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{l-k} P\{N_{d,RT}(i\tau) = l, J_r(i\tau) = l/J_r((n-1)\tau) = m\} & 1 \leq k \leq C \end{cases}$$

Using, this marginal probability and unconditioning on the phases of the input traffic MMPPs at the uplink, we can calculate the exact CLR (by using multinomial coefficients) of real-time traffic, since the arrivals from each uplink are independent. But this process would easily become computationally difficult when N and C increases. Although this will solve obtaining only CLR of real-time traffic but it cannot be helpful to solve for the performance non-real-time queueing process at the downlink. Therefore, next, we consider solving for the joint distribution of $N_{a,RT}(n\tau)$ and $N_{a,RT}((n+j)\tau)$. In other words, we need to know the joint distribution of $N_{d,RT}(n\tau)$ and $N_{d,RT}((n+j)\tau)$. For $j \geq 1$,

$$\sum_{l1}\sum_{l2}\sum_{l3}\sum_{l4} P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2, J_r((n+j-1)\tau) = l3,$$
$$N_{d,RT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4\}$$

$$= \quad P\{N_{d,RT}(n\tau) = i1, N_{d,RT}((n+j)\tau) = i2\}$$

The details of evaluating the joint probability distribution on the left hand side of the above equation are as shown below.

We consider only those random variables on which $N_{d,RT}(n\tau)$ and $N_{d,RT}((n+j)\tau)$ depend. By using the Markov property we can simplify the following joint probability as shown below.

$$P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2, J_r((n+j-1)\tau) = l3,$$
$$N_{d,RT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4\}$$

$$= \quad P\{N_{d,RT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4/N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1,$$
$$J_r(n\tau) = l2, J_r((n+j-1)\tau) = l3\} \times P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1,$$
$$J_r(n\tau) = l2, J_r((n+j-1)\tau) = l3\}$$

where $P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2, J_r((n+j-1)\tau) = l3\}$ can be further simplified as shown below.

$$P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2, J_r((n+j-1)\tau) = l3\}$$

$$= \quad P\{J_r((n+j-1)\tau) = l3/N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2\}$$
$$\times P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2\}$$

where $P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2\}$ can further be simplified as shown below.

$$P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2\} \quad =$$
$$P\{N_{d,RT}(n\tau) = i1, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\} \times P\{J_r((n-1)\tau) = l1\}$$

Based on the Markov property these conditional probabilities, and hence the joint distribution with which we started the simplification. can be written as shown below.

$$P\{N_{d,RT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_r(n\tau) = l2, J_r((n+j-1)\tau) = l3,$$
$$N_{d,RT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4\}$$

$$\begin{aligned} &= P\{N_{d,RT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4/J_r((n+j-1)\tau) = l3\} \\ &\qquad \times P\{J_r((n+j-1)\tau) = l3/J_r(n\tau) = l2\} \\ &\times P\{N_{d,RT}(n\tau) = i1, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\} \times P\{J_r((n-1)\tau) = l1\} \end{aligned}$$

We already obtained the expression for the first factor, and third factor of the right hand side of the above equation. The fourth factor is merely the steady state probability of the phase of the real-time MMPP. The only unknown third factor can be expressed in terms of the transition rate matrix of the phase process of the real-time MMPP, i.e.,

$$P\{J_r((n+j-1)\tau) = l3/J_r(n\tau) = l2\} = \left[e^{Q_r(j-1)\tau}\right]_{l2l3}$$

Once we calculate $P\{N_{d,RT}(n\tau) = i1, N_{d,RT}((n+j)\tau) = i2\}$, the joint probability $P\{N_{a,RT}(n\tau) = i1, N_{a,RT}((n+j)\tau) = i2\}$ can be written as follows.

$$P\{N_{a,RT}(n\tau) = i1, N_{a,RT}((n+j)\tau) = i2\}$$

$$\begin{aligned} = \sum_{m1=i1}^{C} \sum_{m2=i2}^{C} \binom{m1}{i1} \left(\frac{1}{N}\right)^{i1} \left(1-\frac{1}{N}\right)^{m1-i1} \binom{m2}{i2} \left(\frac{1}{N}\right)^{i2} \left(1-\frac{1}{N}\right)^{m2-i2} \\ \times P\{N_{d,RT}(n\tau) = m1, N_{d,RT}((n+j)\tau) = m2\} \end{aligned}$$

Using this joint probability, we can evaluate the covariance function of number of real-time arrivals from a single uplink to a particular downlink, as shown below.

$$\begin{aligned} Cov_{ind}(N_{a,RT}(n\tau), N_{a,RT}((n+j)\tau)) &= \\ E(N_{a,RT}(n\tau) \times N_{a,RT}((n+j)\tau)) - E(N^2_{a,RT}(n\tau)) & \qquad (6.1) \end{aligned}$$

6.3 Covariance Function of Non-real-time Arrival Process to the Downlink from an Uplink

In this section, we describe the method to evaluate the covariance function of non-real-time traffic arriving to a particular downlink from a single uplink, as was done for the case of real-time traffic in the previous section. Let $N_{a,NRT}(i\tau)$ be the number of non-real-time arrivals from the uplink to the particular downlink of interest during the i-th frame. Thus, we need to find the joint distribution of $N_{a,NRT}(n\tau)$ and $N_{a,NRT}((n+j)\tau)$. In other words, we need to know the joint distribution of $N_{d,NRT}(n\tau)$ and $N_{d,NRT}((n+j)\tau)$, where $N_{d,NRT}(n\tau)(N_{d,NRT}((n+j)\tau))$ is the number of non-real-time departures from the uplink during the frame $n\tau((n+j)\tau)$. From this joint distribution of $N_{d,NRT}(n\tau)$ and $N_{d,NRT}((n+j)\tau)$, we can obtain the joint distribution of $N_{a,NRT}(n\tau)$ and $N_{a,NRT}((n+j)\tau)$, as for the case of real-time traffic. From the law of total probability, we have.

$$\begin{aligned}\sum_{l1}\sum_{l2}\sum_{l3}\sum_{l4}\sum_{j1}\sum_{j2}\sum_{j3}\sum_{j4}\sum_{s1}\sum_{s2}\sum_{s3} P\{&N_{d,NRT}(n\tau) = i1, J_r((n-1)\tau) = l1,\\ J_d((n-1)\tau) = j1, q((n-1)\tau) = s1, &J_r(n\tau) = l2, J_d(n\tau) = j2, q(n\tau) = s2,\\ J_r((n+j-1)\tau) = l3, &J_d((n+j-1)\tau) = j3, q((n+j-1)\tau) = s3,\\ N_{d,NRT}((n+j)\tau) = i2, &J_r((n+j)\tau) = l4, J_d((n+j)\tau) = j4\}\\ = \quad P\{&N_{d,NRT}(n\tau) = i1, N_{d,NRT}((n+j)\tau) = i2\}\end{aligned}$$

The details of evaluating the joint probability distribution on the left hand side of the above equation are as shown below. We consider only those random variables on which $N_{d,NRT}(n\tau)$ and $N_{d,NRT}((n+j)\tau)$ depend. Let $N_d(n\tau)$ be the random variable representing number of nrt packet arrivals during n-th frame. By using the Markov property we can simplify the following joint probability as shown below.

$$\begin{aligned}P\{N_{d,NRT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1, q((n-1)\tau) = s1,\\ J_r(n\tau) = l2, J_d(n\tau) = j2, q(n\tau) = s2, J_r((n+j-1)\tau) = l3, J_d((n+j-1)\tau) = j3,\\ q((n+j-1)\tau) = s3, N_{d,NRT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4, J_d((n+j)\tau) = j4\}\end{aligned}$$

$$\begin{aligned} = \quad & P\{N_{d,NRT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4, J_d((n+j)\tau) = j4/N_{d,NRT}(n\tau) = i1, \\ & J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1, q((n-1)\tau) = s1, J_r(n\tau) = l2, J_d(n\tau) = j2. \\ & q(n\tau) = s2. J_r((n+j-1)\tau) = l3. J_d((n+j-1)\tau) = j3. q((n+j-1)\tau) = s3\} \\ & \times P\{N_{d,NRT}(n\tau) = i1. J_r((n-1)\tau) = l1. J_d((n-1)\tau) = j1. q((n-1)\tau) = s1. \\ & J_r(n\tau) = l2. J_d(n\tau) = j2, q(n\tau) = s2. J_r((n+j-1)\tau) = l3. \\ & J_d((n+j-1)\tau) = j3, q((n+j-1)\tau) = s3\} \end{aligned}$$

where the joint probability $P\{N_{d,NRT}(n\tau) = i1. J_r((n-1)\tau) = l1. J_d((n-1)\tau) = j1, q((n-1)\tau) = s1. J_r(n\tau) = l2. J_d(n\tau) = j2, q(n\tau) = s2. J_r((n+j-1)\tau) = l3. J_d((n+j-1)\tau) = j3. q((n+j-1)\tau) = s3\}$ can be simplified as shown below.

$$\begin{aligned} & P\{N_{d,NRT}(n\tau) = i1. J_r((n-1)\tau) = l1. J_d((n-1)\tau) = j1. q((n-1)\tau) = s1. \\ & J_r(n\tau) = l2. J_d(n\tau) = j2. q(n\tau) = s2. J_r((n+j-1)\tau) = l3. \\ & J_d((n+j-1)\tau) = j3. q((n+j-1)\tau) = s3\} \end{aligned}$$

$$\begin{aligned} = \quad & P\{q((n+j-1)\tau) = s3. J_r((n+j-1)\tau) = l3, J_d((n+j-1)\tau) = j3/ \\ & N_{d,NRT}(n\tau) = i1. J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1. q((n-1)\tau) = s1, \\ & J_r(n\tau) = l2. J_d(n\tau) = j2, q(n\tau) = s2\} \\ & \times P\{N_{d,NRT}(n\tau) = i1, J_r((n-1)\tau) = l1. J_d((n-1)\tau) = j1. q((n-1)\tau) = s1, \\ & J_r(n\tau) = l2. J_d(n\tau) = j2. q(n\tau) = s2\} \end{aligned}$$

where $P\{N_{d,NRT}(n\tau) = i1, J_r((n-1)\tau) = l1. J_d((n-1)\tau) = j1. q((n-1)\tau) = s1, J_r(n\tau) = l2. J_d(n\tau) = j2. q(n\tau) = s2\}$ can be simplified as shown below.

$$\begin{aligned} & P\{N_{d,NRT}(n\tau) = i1. J_r((n-1)\tau) = l1. J_d((n-1)\tau) = j1, q((n-1)\tau) = s1, \\ & J_r(n\tau) = l2. J_d(n\tau) = j2, q(n\tau) = s2\} \end{aligned}$$

$$\begin{aligned} = \quad & P\{N_{d,NRT}(n\tau) = i1, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/ \\ & q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \\ & \times P\{q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \end{aligned}$$

Based on the Markov property these conditional probabilities, and hence the joint distribution with which we started the simplification, can be written as shown below.

$$\begin{aligned} & P\{N_{d,NRT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1, q((n-1)\tau) = s1, \\ & J_r(n\tau) = l2, J_d(n\tau) = j2, q(n\tau) = s2, J_r((n+j-1)\tau) = l3, J_d((n+j-1)\tau) = j3, \\ & q((n+j-1)\tau) = s3, N_{d,NRT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4, J_d((n+j)\tau) = j4\} \end{aligned}$$

$$\begin{aligned} = \quad & P\{N_{d,NRT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4, J_d((n+j)\tau) = j4/ \\ & q((n+j-1)\tau) = s3, J_r((n+j-1)\tau) = l3, J_d((n+j-1)\tau) = j3\} \\ & \times P\{q((n+j-1)\tau) = s3, J_r((n+j-1)\tau) = l3, J_d((n+j-1)\tau) = j3/ \\ & q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2\} \\ & \times P\{N_{d,NRT}(n\tau) = i1, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/ \\ & q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \\ & \times P\{q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \end{aligned}$$

The main focus here, is to derive the conditional probabilities, the third factor on the right hand side of the above equation. The first factor can be easily obtained from the third factor. The second factor can be obtained by using the transition probability matrix of the non-real-time queue at the uplink. The fourth factor is the steady state joint probability of non-real-time queue at the uplink and the phases of real-time and non-real-time MMPPs.

We proceed to derive the expression for $P\{N_{d,NRT}(n\tau) = k, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$.

For $s1 = 0$,

For $k = 0$,

For $s2 = 0$,

$$\begin{aligned} & P\{N_{d,NRT}(n\tau) = 0, q(n\tau) = 0, J_r(n\tau) = l2, J_d(n\tau) = j2/ \\ & \quad q((n-1)\tau) = 0, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \\ & = \quad P\{N_d(n\tau) = 0, J_d(n\tau) = j2/J_d((n-1)\tau) = j1\} \\ & \qquad \times P\{J_r(n\tau) = l2/J_r((n-1)\tau) = l1\} \end{aligned}$$

For $s2 \geq 1$,

$$\begin{aligned} & P\{N_{d,NRT}(n\tau) = 0, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/ \\ & \quad q((n-1)\tau) = 0, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \\ & = \quad P\{N_d(n\tau) = s2, J_d(n\tau) = j2/J_d((n-1)\tau) = j1\} \\ & \qquad \times P\{N_r(n\tau) \geq C, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\} \end{aligned}$$

For $1 \leq k \leq (C-1)$,

For $s2 = 0$,

$$\begin{aligned} & P\{N_{d,NRT}(n\tau) = k, q(n\tau) = 0, J_r(n\tau) = l2, J_d(n\tau) = j2/ \\ & \quad q((n-1)\tau) = 0, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \\ & = \quad \sum_{i=0}^{C-k} (P\{N_d(n\tau) = k, J_d(n\tau) = j2/J_d((n-1)\tau) = j1\} \\ & \qquad \times P\{N_r(n\tau) = i, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}) \end{aligned}$$

For $s2 \geq 1$,

$$P\{N_{d,NRT}(n\tau) = k, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = 0, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \quad P\{N_r(n\tau) = (C-k), J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$
$$\times P\{N_d(n\tau) = (s2+k), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$

For $k = C$,

$$P\{N_{d,NRT}(n\tau) = k, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = 0, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \quad P\{N_r(n\tau) = 0, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$
$$\times P\{N_d(n\tau) = (s2+C), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$

For $s1 \geq 1$,
For $k = 0$,
For $s2 \geq s1$,

$$P\{N_{d,NRT}(n\tau) = 0, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \quad P\{N_d(n\tau) = (s2-s1), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$
$$\times P\{N_r(n\tau) \geq C, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$

For $s2 < s1$,

$$P\{N_{d,NRT}(n\tau) = 0, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \quad = \quad 0$$

For $1 \leq k \leq (C-1)$,

For $1 \leq s1 \leq k$,

For $s2 = 0$,

$$P\{N_{d,NRT}(n\tau) = k, q(n\tau) = 0, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \sum_{i=0}^{C-k} P\{N_d(n\tau) = (k-s1), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$
$$\times P\{N_r(n\tau) = i, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$

For $s2 \geq 1$,

$$P\{N_{d,NRT}(n\tau) = k, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \quad P\{N_r(n\tau) = (C-k), J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$
$$\times P\{N_d(n\tau) = s2 + (k-s1), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$

For $s1 > k$,

For $s2 \geq (s1-k)$,

$$P\{N_{d,NRT}(n\tau) = k, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \quad P\{N_r(n\tau) = (C-k), J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$
$$\times P\{N_d(n\tau) = s2 - (s1-k), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$

For $s2 < (s1-k)$,

$$P\{N_{d,NRT}(n\tau) = k, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \quad = \quad 0$$

For $k = C$,

For $1 \leq s1 \leq C$,

For $s2 \geq 0$,

$$P\{N_{d.NRT}(n\tau) = C, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \quad P\{N_r(n\tau) = 0, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$
$$\times P\{N_d(n\tau) = s2 + (C - s1), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$

For $s1 > C$,

For $s2 \geq (s1 - C)$,

$$P\{N_{d.NRT}(n\tau) = C, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\}$$

$$= \quad P\{N_r(n\tau) = 0, J_r(n\tau) = l2/J_r((n-1)\tau) = l1\}$$
$$\times P\{N_d(n\tau) = s2 - (s1 - C), J_d(n\tau) = j2/J_d((n-1)\tau) = j1\}$$

For $s2 < (s1 - C)$,

$$P\{N_{d.NRT}(n\tau) = C, q(n\tau) = s2, J_r(n\tau) = l2, J_d(n\tau) = j2/$$
$$q((n-1)\tau) = s1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1\} \quad = \quad 0$$

With these conditional probabilities evaluated, we can express the joint probability $P\{N_{d.NRT}(n\tau) = i1, J_r((n-1)\tau) = l1, J_d((n-1)\tau) = j1, q((n-1)\tau) = s1, J_r(n\tau) = l2, J_d(n\tau) = j2, q(n\tau) = s2, J_r((n+j-1)\tau) = l3, J_d((n+j-1)\tau) = j3, q((n+j-1)\tau) = s3, N_{d.NRT}((n+j)\tau) = i2, J_r((n+j)\tau) = l4, J_d((n+j)\tau) = j4\}$ in known terms and hence the joint probability $P\{N_{d.NRT}(n\tau) = i1, N_{d.NRT}((n+j)\tau) = i2\}$ can be evaluated.

Therefore, we can evaluate the joint probability $P\{N_{a,NRT}(n\tau) = i1, N_{a,NRT}((n+j)\tau) = i2\}$ by using the following relationship (similar to the real-time case). Note that to compute this joint probability we need the joint steady state probability distribution of uplink queue size, phases of real-time and non-real-time traffic, and the method to calculate this, which was presented in Sections 5.2 and 5.3.

$$P\{N_{a,NRT}(n\tau) = i1, N_{a,NRT}((n+j)\tau) = i2\} \;=$$

$$\sum_{m1=i1}^{C} \sum_{m2=i2}^{C} \binom{m1}{i1} \left(\frac{1}{N}\right)^{i1} \left(1 - \frac{1}{N}\right)^{m1-i1} \binom{m2}{i2} \left(\frac{1}{N}\right)^{i2} \left(1 - \frac{1}{N}\right)^{m2-i2}$$
$$P\{N_{d,NRT}(n\tau) = m1, N_{d,NRT}((n+j)\tau) = m2\}$$

Therefore, we can evaluate the covariance of the non-real-time arrival process, as shown below.

$$Cov_{ind}(N_{a,NRT}(n\tau), N_{a,NRT}((n+j)\tau))$$

$$= \; E(N_{a,NRT}(n\tau) \times N_{a,NRT}((n+j)\tau)) - E(N_{a,NRT}^{2}(n\tau)) \tag{6.2}$$

As we discussed for the real-time case, the departure process $(N_{d,NRT}(n\tau))$ is actually a discrete-time Markov modulated process with phase transitions occuring according to the phase of the non-real-time MMPP and non-real-time queue status with the batch size, which varies over 0 to C with the dependence on the phase of non-real-time MMPP and uplink queue status. However, characterising this as a discrete-time Markov modulated process would mean taking into account an infinite queue and hence an infinite phase, which would result in an infeasible computation or analysis. With the analysis presented here, we can go one step further to calculate any n-th order statistics of the counting process as and when the matching techniques need because of the Markov property.

The last and final step would involve matching the second-order and first-order statistics to the corresponding 2-state MMPP and solving for four parameters of 2-state MMPP. The covariance of net arrival process to a particular downlink is the sum of covariance of arrival process to the downlink from each uplink. We choose the covariance at two different points and mean and variance of number of net arrival process to match to the corresponding values of 2-state MMPP[84]. The matching algorithm that we used for non-real-time traffic case is presented as shown below. The covariance of 2-state MMPP with parameters $\lambda_1, \lambda_2, \sigma_1, \sigma_2$ is given by[78].

$$Cov(N(n\tau), N((n+j)\tau)) = \frac{\sigma_1\sigma_2 \cdot (1 - e^{-(\sigma_1+\sigma_2)\tau})^2 \cdot (\lambda_1 - \lambda_2)^2 e^{-(\sigma_1+\sigma_2)j\tau}}{(\sigma_1+\sigma_2)^4} \tag{6.3}$$

From Equation 6.2.

$$Cov_{tot}(N_{NRT}(n\tau), N_{NRT}((n+1)\tau)) = N * Cov_{ind}(N_{a,NRT}(n\tau), N_{a,NRT}((n+1)\tau))$$

$$Cov_{tot}(N_{NRT}(n\tau), N_{NRT}((n+j)\tau)) = N * Cov_{ind}(N_{a,NRT}(n\tau), N_{a,NRT}((n+j)\tau))$$

where N represents the total number of uplinks. From Equation 6.3 and using the above relationships. we can write.

$$r = ln(Cov_{tot}(N_{NRT}(n\tau), N_{NRT}((n+1)\tau))/Cov_{tot}(N_{NRT}(n\tau), N_{NRT}((n+j)\tau))) \times \frac{1}{(j-1)\tau}$$

where $r = \sigma_1 + \sigma_2$. We choose $\sigma_1 = r/2$. From Equation 6.3.

$$(\lambda_1 - \lambda_2)^2 = \frac{Cov_{tot}(N_{NRT}(n\tau), N_{NRT}((n+j)\tau)) * r^4}{\sigma_1\sigma_2(1 - e^{-r\tau})^2 e^{-rj\tau}} \tag{6.4}$$

$$\frac{\lambda_1\sigma_2 + \lambda_2\sigma_1}{r} = \frac{N * E(N_{a,NRT}(n\tau)}{\tau} \tag{6.5}$$

Using Equations 6.4 and 6.5. we solve for λ_1 and λ_2.

The matching algorithm for real-time traffic case is the same as above. except that we use corresponding real-time traffic parameters.

6.4 Illustrative Results

The CLR of rt-VBR real-time traffic and stationary distribution of nrt-VBR queue length at the downlink can be derived by using the analysis presented in the previous section. First. using the above analysis. we compute the covariance of the net arrival process of real-time traffic and non-real-time traffic at the downlink. We assume symmetric and independent traffic conditions at the inputs. Also. we assume uniform

Traffic Type	Data Dominant
RT Voice	20%
RT Video	10%
NRT Data	70%

Table 6.1: The nature of traffic mix.

destination distribution. Therefore the covariance of the net arrival process will simply correspond to the sum of the covariance of arrival processes from all uplinks to the particular downlink. The same conditions hold for mean and variance of net arrival process at the downlink. In the following, we present some examples with illustrative results[94][96]. We take into account the overhead associated with the header information and assume a WATM cell size of 54 bytes to carry an ATM SDU of 48 bytes. The WATM cell contains a header of 6 bytes. An additional byte as compared to the standard 53-byte ATM cell is for the data link error control mechanism[26]. A TDMA frame of 1msec is assumed. For the illustrative examples, we consider uplink capacity of $C_u = 5Mbps$. With a 1 msec frame length, this rate yields $C = 11$ slots per frame. We consider a particular mix of traffic with data traffic being dominant. The salient characteristics of voice, video and data traffic are shown in Table 5.1. The mix of the traffic types, for the "data dominant" case is shown in Table 6.1.

In Figure 6.1 and Figure 6.2, we show both the simulation and approximation results of logarithm of survivor function of non-real-time queue at the downstream link for the case of 50% and 80% total load respectively. For our simulations, we choose a confidence level of 95%; all of the simulation points are within the confidence interval of 5% with this confidence level. For the details of calculating the confidence interval for occupancy probability, refer Section 4.1. The CLR of real-time traffic at the output queue of the ATM switch is 8.761×10^{-8} obtained through approximation and 1.703×10^{-7} obtained through simulation, for the case of load of

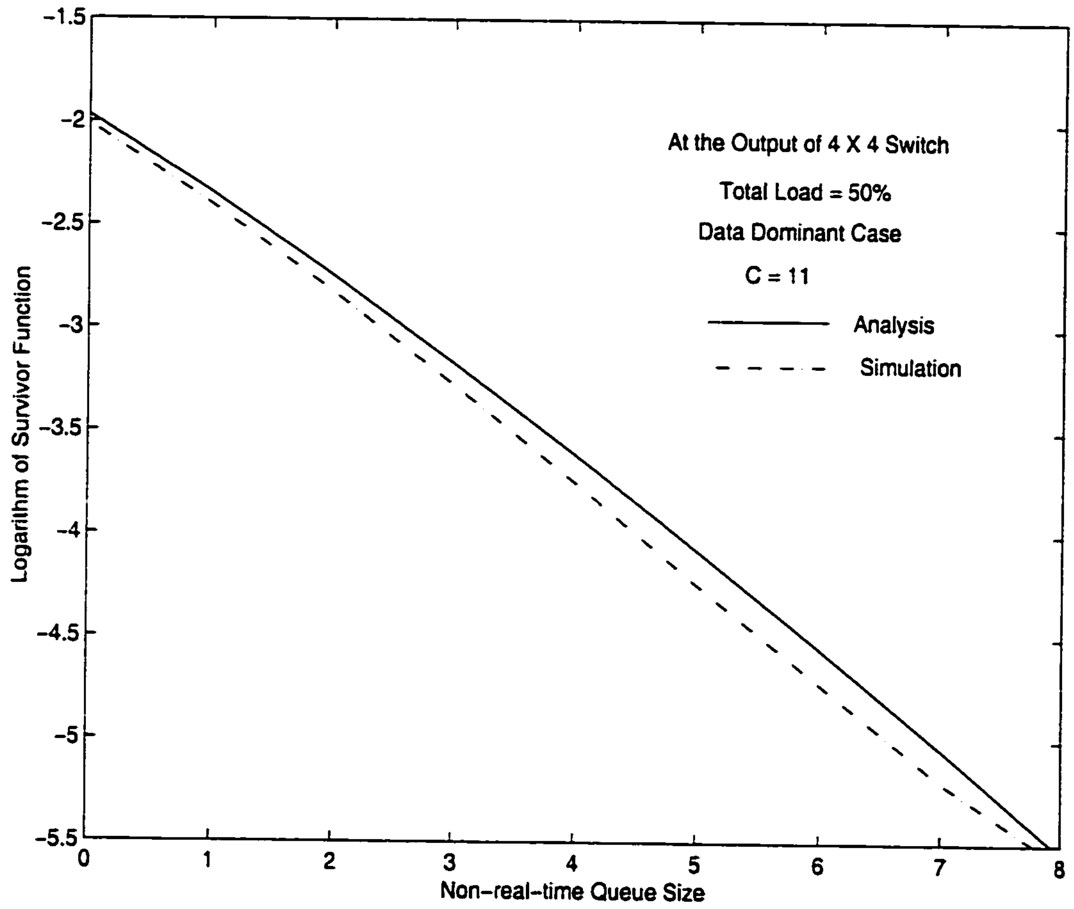

Figure 6.1: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for $C_u = 5Mbps$ and for a load 50%.

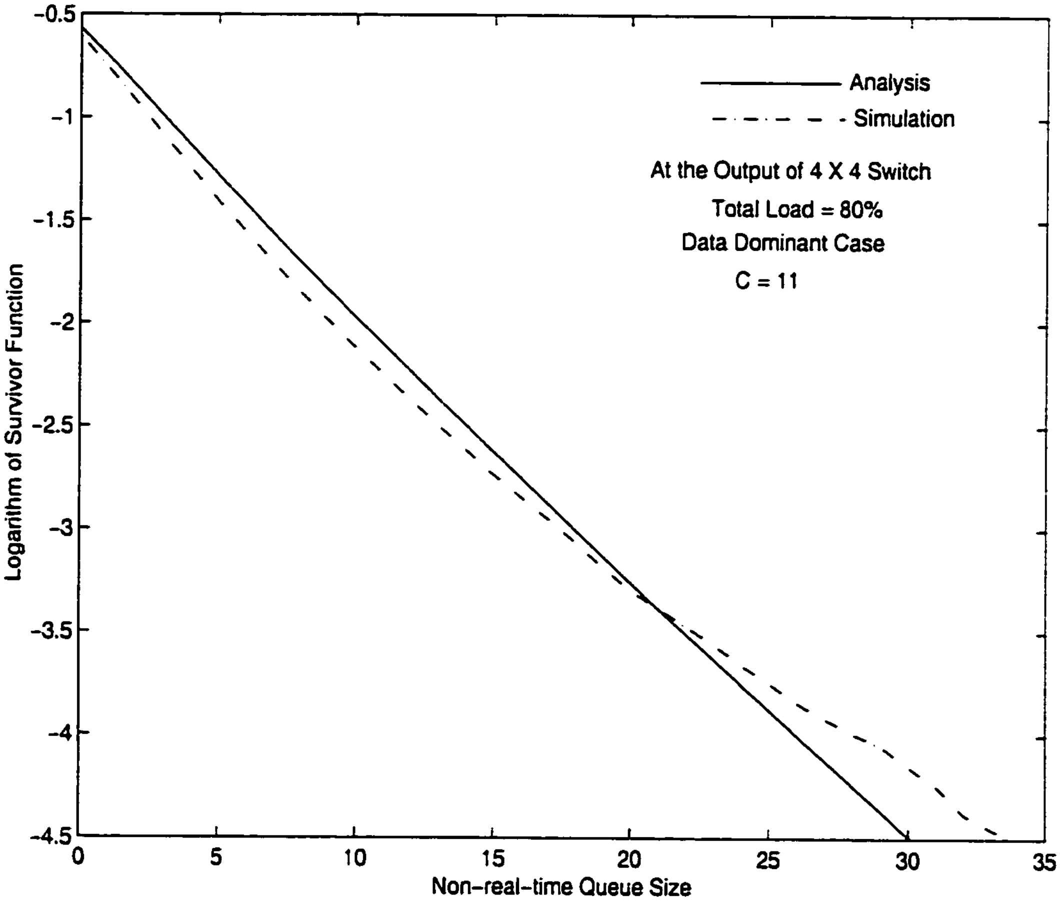

Figure 6.2: The logarithm of the survivor function of NRT data queue length vs. NRT data queue length for $C_u = 5Mbps$ and for a load 80%.

50%, and 1.8567×10^{-5} obtained through approximation and 9.856×10^{-6} obtained through simulation, for the case of load of 80%. We have used our analysis on the arrival process and approximation to 2-state MMPPs for both real-time traffic and non-real-time traffic, as mentioned in the previous section. Note that the analysis of covariance of arrival process is exact. However, in performing computation of covariance function of non-real-time traffic, we have to truncate the infinite matrix. If the truncation can be performed such that the truncated rows are close to zero, then the computation of covariance can be very close to actual value. We proceeded in computing the rows whose probability values are greater than 10^{-250}(in Matlab). The platform on which the computations were run, was SUN Ultra workstations. The other approximation needed is matching these characteristics to MMPP's characteristics, which has extensively been discussed in the literature[84][69][71][72][73][74]. We see clearly that our approximation results are very close to the actual results.

In the next chapter we present our conclusions and future work to be carried from our research.

Chapter 7

Conclusions and Future Work

In this thesis, we developed analytical modeling techniques that can be used to evaluate the QoS of RT and NRT traffic with priority based scheduling at the nodes in a differentiated services network. The RT and NRT traffic arriving at each node are modeled as MMPPs. We presented typical system configurations in terrestrial wireless and satellite wireless networks, where this methodology is particularly applicable. The solution methodology in evaluating the CLR performance of RT and NRT traffic is based on the matrix-geometric technique, which is computationally intensive but much more stable than other techniques in providing solutions.

Since priority is given to RT traffic over NRT traffic, the performance of RT traffic can be analysed independently. We obtained an expression for calculating CLR of RT traffic at the scheduler. On the other hand, since handling NRT traffic, depends on how many RT cells arrive during the previous frame, an embedded Markov chain analysis of NRT queue length at the beginning of frame periods after scheduling RT and NRT traffic is developed by conditioning on number of RT cell arrivals, NRT cell arrivals, and phases of RT MMPP and NRT MMPP. By identifying the elements of the transition probability matrix with the structure of BMAP/G/1, we show how the steady-state occupancy probability of NRT queue can be obtained. We present results for the case of TDMA frame duration of 1 millisecond with C = 11 and C = 28 slots per frame for different mixes of RT and NRT traffic and examine the statistical multiplexing gain for all these cases. Each slot is assumed to carry a payload of 48 bytes. These cases correspond to the uplink capacity of 5 Mbps and 12.5 Mpbs in a typical terrestrial wireless cluster. Our results can be used in deciding the type and number of RT and NRT connections that can be allowed at the user terminals by the base station scheduler and setting the NRT queue size for a particular CLR. To compare the performance with actual on-off source models against approximating MMPP models, we presented both analytical and simulation results corresponding to these cases. We also present results for the case of TDMA frame duration of 24 milliseconds with C = 64 slots per frame for different mixes of RT and NRT traffic and examine the statistical multiplexing gain for all these cases. These types of parameters are particularly typical of satellite networks.

To evaluate performance of priority based scheduling at any downstream nodes, first of all we model the arrival process of RT traffic and NRT traffic as approximating MMPPs. To do this, we derive the covariance function of RT arrivals/NRT arrivals from an input port of the switch to an output port by using the Markov property, assuming the priority based scheduling with MMPP sources at the input port and each arriving RT/NRT fixed size packets or cells choosing an output port according to the uniform distribution. Although this covariance function can be easily computed for RT traffic, the computation of covariance function of NRT traffic is much more intensive. Using the covariance function values at two different lags, and mean and variance of arrival rate of RT and NRT traffic, we obtain the parameters of the approximating 2-state MMPP for RT traffic and the parameters of the approximating 2-state MMPP for NRT traffic. With these MMPP representations, we use the analysis outlined in the previous paragraph, to evaluate the CLR performance of RT and NRT traffic at the output port with the priority based scheduling. We present both analytical and simulation results for the case of TDMA frame duration of 1 millisecond with C = 11 slots per frame for the case of 4X4 switch at an output port for the data dominant case at two different loads of 50% and 80%. The CLR values of

RT traffic obtained through analysis and simulation are in good agreement. Similarly, the survivor functions of NRT traffic obtained through analysis and simulation are also in good agreement.

Thus, our research can be used to analyse the performance of priority based scheduling at any node in the network. If the end-to-end performance needs to be computed, we can do so by using the performance quantities obtained at each node and using an independence assumption over the links.

In performing the computation with larger capacities and MMPPs with many more states, the convergence of matrix-geometric solution is rather slow both because of higher processing time and storage complexity. To overcome this problem, mathematical research on better techniques is going on[88]. Thus, the future work could include discovering new fast techniques or improving existing techniques for obtaining matrix geometric solution and/or obtaining approximate estimates of survivor function(tail probabilities) by analytical approximations. Also, future work could include making the covariance function computation less intensive in terms of time and memory by establishing a structural computation through theoretical analysis.

Appendix A

Simulation Model of an Ideal Switch

A.1 Outline of the Simulation Model

The node module used in this simulation model is shown in figure A.1.

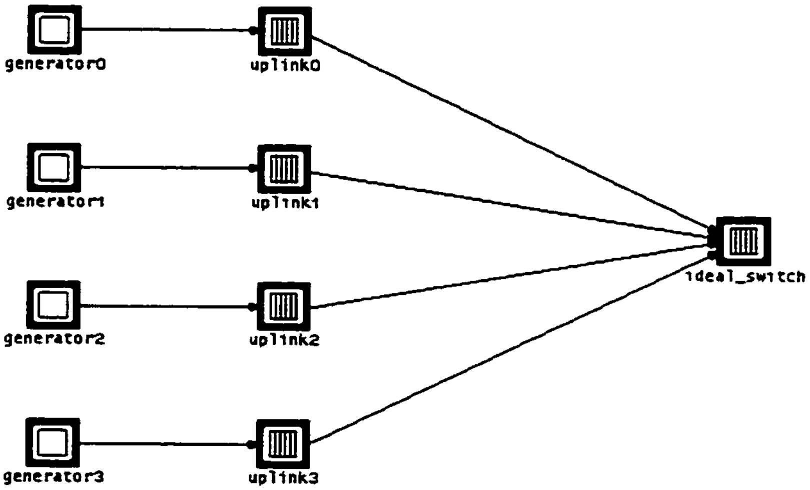

Figure A.1: The node module of the Opnet simulation model of the Ideal Switch.

The Generator node module, would be generating packets according to the process model called "trf_generator". The "trf_generator" process model would in turn invoke the process models of voice, video and data MMPP packet generators.

At the uplink module the following activities are implemented:

- queuing real-time packets and data packets in two separate queues as and when they arrive;
- at the end of every frame period, the real-time packets are to be transmitted on a priority basis. If there are real-time packets that cannot be served they have to be simply discarded. If there is capacity after sending real-time packets, the corresponding number of data packets will be sent from the data queue.

At the Switch, the following activities are implemented:

- Since we are considering the symmetric case, packets not destined to a particular downlink(we consider only one downlink) will be destroyed. At the switch, only those packets with one particular destination will be stored for the downstream link.
- Thus, at the ideal_switch module, we also implement the similar queueing activities as those at the uplink module, by using similar process modules.
- Finally the module also collects the statistics on survivor function of non-real-time queue at the downstream link and the CLR of real-time arrivals.

A.1.1 Process Models

At the uplink, the two activities mentioned above for the uplink processor module are implemented by a process model called "input_buffer". Thus, there is a process model that simply receives packets and queues them in either real-time buffer or data queue depending on the type of the packet, each time an arrival occurs from the generator module. The corresponding state diagram of the root process model named "input_buffer" is shown in figure A.2. The same process model at the end of every frame period transmits real-time packets in a priority manner over non-real-time packets.

At the switch module named "ideal_switch". the process model named "ideal_switch" waits for the packets to be received from the input streams (which are four). Once packets arrive. it simply chooses those packets whose destination is "0" and destroys all other packets. From the chosen packets. it extracts the type of the packet. If it is a data packet, it queues them in the non-real-time data queue. and if it is real-time packet, it stores them in the real-time buffer. At the end of every frame period, the same priority based scheduling operates on the real-time buffer and the non-real-time queue. The CLR of real-time data and survivor function of non-real-time queue are calculated as earlier. The corresponding state diagram of the "ideal_switch" process model is similar to that shown in figure A.2. Note that although the state diagram of "input_buffer" and "ideal_switch" looks similar the code is different for "ideal_switch" since it deals with only those packets whose destination is "0".

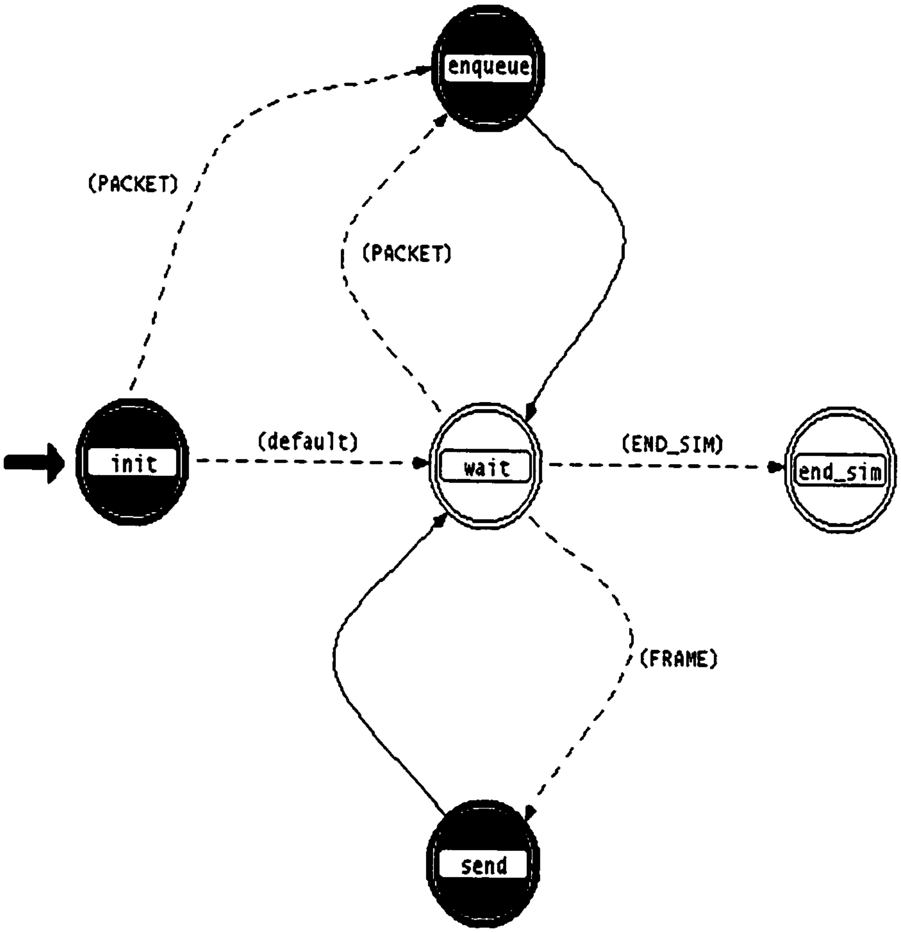

Figure A.2: The state diagram of input_buffer process model.

Appendix B

Simulation Model of a Scheduler

B.1 Outline of the Simulation Model

The node module used in this simulation model is shown in figure B.1.

At the terminal module the following activities are implemented:

- queuing real-time packets and data packets in two separate queues as and when they arrive;
- at the end of every frame period, the request packet of number of real-time packets in the buffer and number of non-real-time packets in the non-real-time data queue, are to be transmitted to the scheduler.
- After receiving the real-time and non-real-time grants from scheduler, cell loss ratio is calculated for real-time traffic and real-time queue is flushed out. Upto the non-real-time grants, the packets from non-real-time queue are removed and destroyed. The statistics on non-real-time queue are updated to calculate the survivor function.

At the scheduler module the following activities are implemented:

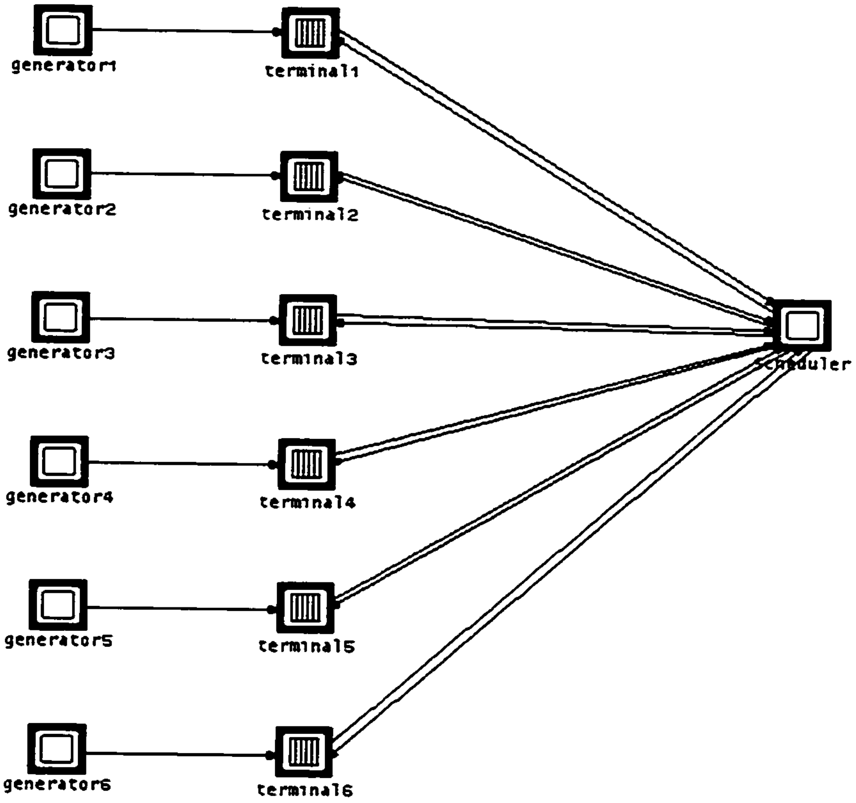

Figure B.1: The node module of the Opnet simulation model of Scheduler.

- Based on the received requests, the scheduler distributes the Capacity in a round robin manner until either the capacity expires or all real-time requests are satisfied first.

- If there is still capacity available, the scheduler distributes the remaining capacity in a round robin manner until either the capacity expires or all non-real-time requests are satisfied.

- It updates the next terminal to be served first in the next cycle. This terminal update is done to consider the requests from terminals in a round robin manner apart from allocating in a round robin manner.

- It calculates the aggregate CLR for real-time traffic of all terminals and survivor function of aggregate non-real-time queue of all terminals.

- It sends the grants to the terminals.

B.1.1 Process Models

At the terminals. the activities mentioned above are implemented by the process model called "terminal_input_buffer" shown in Figure B.2. The activities have one to one correspondence with the states as shown in Figure B.2.

At the scheduler the process module named "base_station_scheduler", waits for the request packets to be received from the input streams every frame and performs the activities as mentioned in the list of activities at the scheduler. The state diagram of this process module is shown in Figure B.3. For proper synchronisation of the terminal process module and scheduler process module the frames start a bit later at the scheduler as compared to the start times at the terminals.

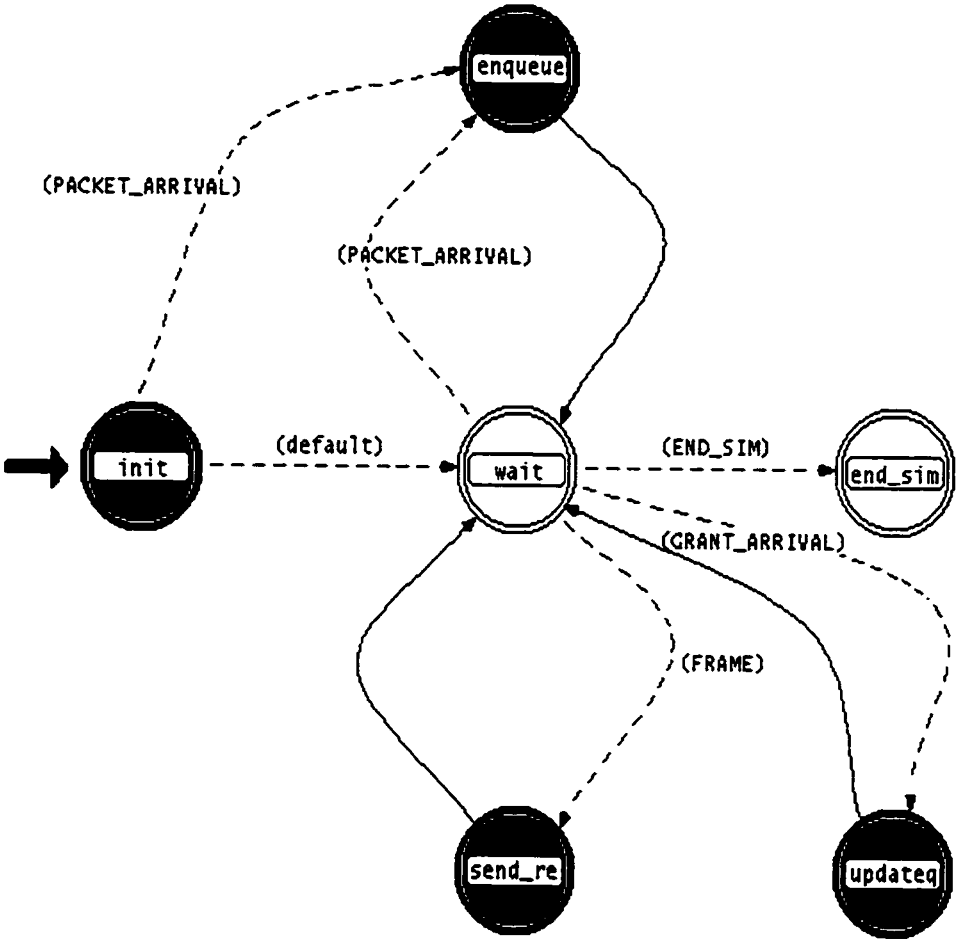

Figure B.2: The state diagram of terminal_input_buffer process model.

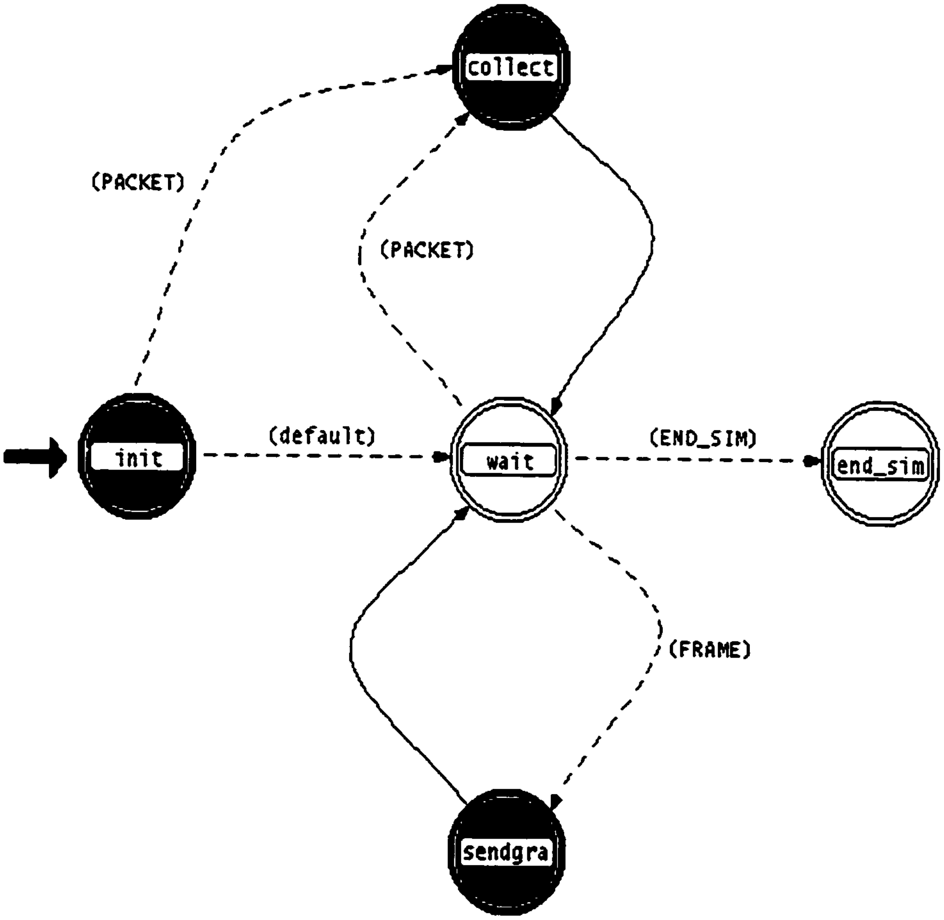

Figure B.3: The state diagram of base_station_scheduler process model.

Bibliography

[1] P. Takats, S. Irani. "System Architecture for an Advanced Canadian Communications Satellite Demonstration Mission". American Institute of Aeronautics and Astronautics. Washington. March 22-26. 1992.

[2] P. Garland. Tho Le-Ngoc. P. Takats. "Fast Packet Switches for Next Generation Satcom Applications". International Conference on Digital Satellite Communications-9. Copenhagen. May 18-22.1992.

[3] T. Inukai. F. Faris. D-J. Shyy. "On-Board Processing Satellite Network Architectures for Broadband ISDN". American Institute of Aeronautics and Astronautics. Washington. March 22-26. 1992.

[4] P. Garland, Tho Le-Ngoc, U. Mossinger. "Hybrid On-Board Switching Architectures for the Provision of both Circuit and Packet Switched Satellite Services". American Institute of Aeronautics and Astronautics. San Diego. CA. May 3-6. 1994.

[5] M. Bever et. al., "Fast-packet vs. Circuit Switch and Bent Pipe Satellite Network Architecture", International Journal of Satellite Communications. Vol. 17. Numbers 2 and 3, March-June 1999, pp: 83-105.

[6] Enrico Del Re and R. Fantacci. "An Advanced Satellite Communication System with On-board Fast Packet Switching Capabilities". International Journal of Satellite Communications. Vol. 12, 1994, pp: 147-155.

[7] T. Ors, "ATM over Satellite", *http:// www.ee.surrey.ac.uk/ Personal/ T.Ors/ atmsat*, date of update: 17 Dec. 2000.

[8] T. Ors et. al., "A Meshed VSAT Satellite Network Architecture using an On-board ATM Switch". IEEE International Conference on Performance, Computing and Communication, Tempe/Phoenix, Arizona, February 1997.

[9] A. Baiocchi et. al., "Modeling and Dimensioning of an Integrated Circuit and Packet Switching Scheme On-board a Processing Satellite". International Conference on Communications. June 23-27, Dallas, Texas, 1996 pp: 936-941.

[10] H. Koraitim and S. Toheme. "Resource Allocation and Connection Admission Control in Satellite Networks". IEEE Journal on Selected Areas in Communications, Vol. 17, No. 2, February 1999, pp: 360-372.

[11] Tho Le-Ngoc, Tien Hy Bui, and M. Hachicha, "Performance of a Knockout Switch for Multimedia Satellite Communications". 3-8 November 1997, Phoenix, Arizona, Globecom'97.

[12] Whole issue of IEEE Personal Communications, Vol. 4, No. 4, August 1997.

[13] P. Chitre et.al., "Next-Generation Satellite Networks: Architectures and Implementation", IEEE Communications Magazine, Vol. 37, No. 3, March 1999, pp: 30-36.

[14] W. D. Ivanic et. al., "A Network Architecture for a Geostationary Communication Satellite". IEEE Communications Magazine, Vol. 32, No. 7, July 1994, pp: 72-84.

[15] G. Losquadro, "EUROSKYWAY: Satellite System for Interactive Multimedia Services". Proc. Ka-Band Util. Conf., Florence, Italy, Sept. 1996, pp: 13-20.

[16] I. Mertzanis et. al., "Protocol Architectures for Satellite ATM Broadband Networks", IEEE Communications Magazine, Vol. 37, No. 3, March 1999, pp: 46-54.

[17] Application of Teledesic Corporation for a Low-Earth-Orbit Satellite System in the Fixed Satellite Service, (FCC filing), March 21, 1994.

[18] Application of Hughes Communications Galaxy, Inc. before the FCC for two Ka-band, pp: 906-909 Domestic Fixed Communications Satellites , (FCC filing), December 3, 1993.

[19] Tho Le-Ngoc and S. Krishnamurthy, "Performance of Combined Free/Demand Assignment Multiple-Access Schemes in Satellite Communications", International Journal on Satellite Communications, Vol. 14, pp: 11-21, 1996.

[20] Hassan Peyravi, "Medium Access Control Protocols Performance in Satellite Communications", IEEE Communications Magazine, Vol. 37, No. 3, March 1999, pp: 62-71.

[21] G. D. Gordon and W. L. Morgan, "Principles of Communications Satellites", John Wiley and Sons Inc., 1993.

[22] D. Roddy, "Satellite Communications", McGraw Hill, 2nd Edition, 1996.

[23] WAND Project Public Deliverable 1D3, "Report: WAND Design Requirements", *http://www.tik.ee.ethz.ch/w̃and/DOCUMENTS/documents-frame.html*, 30-09-1996.

[24] WAND Project Public Deliverable 3D1, "Report: Wireless ATM MAC Overall Description", *http://www.tik.ee.ethz.ch/w̃and/DOCUMENTS/documents-frame.html*, 31-12-1996.

[25] D. Raychauduri, "Wireless ATM: An Enabling Technology for Multimedia Personal Communication", Wireless Networks, Vol. 2, 1996 pp:163-171.

[26] D. Raychaudhuri et. al., "WATMnet: A Prototype Wireless ATM System for Multimedia Personal Communication", IEEE Journal on Selected Areas in Communications, Vol.15, No. 1, Jan. 1997, pp:83-95.

[27] A. S. Tanenbaum, "*Computer Networks*", Englewood Cliffs, NJ: Prentice-Hall, 1996.

[28] L. Kaufman, "Matrix Methods for Queueing Problems", SIAM Journal on Scientific and Statistical Computing, Vol. 4, No. 3, Sept. 1983, pp: 525-552.

[29] J. Y. Le Boudec, "An Efficient Solution Method for Markov Models of ATM Links with Loss Priorities", IEEE Journal on Selected Areas in Communications, Vol. 9, No. 3, April, 1991, pp: 408-417.

[30] J. Y. Le Boudec, "A generalisation of matrix-geometric solutions for Markov chains", IBM Res. Rep. RZ 4629, 1989.

[31] G. Latouche et. al., "Finite Markov Chains Skip Free in One Direction", Naval Res. Logistic. Quart., Vol. 31, 1984, pp: 571-588.

[32] T. Takine et. al., "Cell Loss and Output Process Analyses of a Finite-Buffer Discrete-Time ATM Queueing System with Correlated Arrivals", IEEE Transactions on Communications, Vol. 43, No. 2/3/4, Feb./Mar./Apr. 1995, pp: 1022-1037.

[33] N. Akar et. al., "TELPACK: An Advanced Teletraffic Analysis Package", IEEE Communications Magazine, Vol. 36, No. 8, pp: 84-87, Aug. 1998.

[34] W. K. Grassman et. al., "Equilibrium Distribution of Block-Structured Markov Chains with Repeating Rows", Journal of Applied Probability, 27, pp: 557-576, 1990.

[35] S. Q. Li and H. D. Sheng, "Discrete Queueing Analysis of Multi-Media Traffic with Diversity of Correlation and Burstiness Properties", Infocom, Bal Harbour. Florida, April 7-11, pp: 368-381, 1991.

[36] J. Ye and S. Q. Li, "Folding Algorithm: A Computational Method for Finite QBD Processes with Level-Dependent Transitions", IEEE Transactions on Communications, Vol. 42, No. 2/3/4, Feb./Mar./Apr. 1994, pp: 625-639.

[37] M. F. Neuts, "**Matrix-Geometric Solutions in Stochastic Models - An Algorithmic Approach**", The John Hopkins University Press. 1981.

[38] Jeremiah F. Hayes. "**Performance Modeling and Analysis of Computer Communication Systems**". Plenum Press, New York. 1984.

[39] Jun Huang and Jeremiah F.Hayes. "A study of BMAP/SMSP(SMVP)/1 Queueing System". International Symposium on Operations Research with Applications in Engg., Technology and Management (ISORA), Beijing, China. Aug. 19-22. 1995.

[40] L. Kleinrock. "**Queueing Systems**". Vol. I. John Wiley and Sons. 1975.

[41] K. Sriram et. al., "Discrete-Time Analysis of Integrated Voice/Data Multiplexers With and Without Speech Activity Detectors", IEEE Journal on Selected Areas in Communications, SAC-1, No. 5. Dec. 1983, pp: 1124-1132.

[42] S. Q. Li and J. W. Mark, "Performance of Voice/Data Integration on a TDM System". IEEE Transactions on Communications, Vol. 33, No. 12, Dec. 1985. pp: 1265-1273.

[43] S. Q. Li, "A General Solution Technique for Discrete Queueing Analysis of Multimedia Traffic on ATM". IEEE Transactions on Communications. Vol. 39. No. 7, July 1991, pp: 1115-1132.

[44] S. Q. Li, "Generating Function Approach for Discrete Queueing Analysis with Decomposable Arrival and Service Markov Chains", Commun. Statist.-Stochastic Models, 9 (3), 1993, pp: 401-420.

[45] H. Bruneel, "Queueing Behaviour of Statistical Multiplexers with Correlated Inputs", IEEE Transactions on Communications, Vol. 36, No. 12, Dec. 1988, pp: 1339-1441.

[46] M. M. Asrin and F. Kamoun, "A Transient Discrete-time Queueing Analysis of the ATM Multiplexer", Performance Evaluation, Vol. 32, 1998, pp: 153-183.

[47] S. Q. Li et. al., "SMAQ: A Measurement-Based Tool for Traffic Modeling and Queueing Analysis Part I: Design Methodologies and Software Architecture", IEEE Communications Magazine, Vol. 36, No. 8, pp: 56-65, Aug. 1998.

[48] D. M. Lucantoni et. al., "Methods for Performance Evaluation of VBR Video Traffic Models", IEEE/ACM Transactions on Networking, Vol. 2, No. 2, pp: 176-180, April, 1994.

[49] D. P. Heyman and T. V. Lakshman, "Source Models for VBR Broadcast-Video Traffic", IEEE/ACM Transactions on Networking, Vol. 4, No. 1, pp: 40-48, Feb. 1996.

[50] D. P. Heyman, "The GBAR Source Model for VBR Videoconferences", IEEE/ACM Transactions on Networking, Vol. 5, No. 4, pp:554-560, August 1997.

[51] A. T. Andersen and B. F. Nielsen, "A Markovian Approach for Modeling Packet Traffic with Long-Range Dependence", IEEE Journal on Selected Areas in Communications, Vol. SAC-16, No. 5, pp:719-732, June 1998.

[52] C. Blondia and O. Casals, "Performance Analysis of Statistical Multiplexing of VBR Sources", Proc. of IEEE Infocom 92, Florence, Italy, pp:828-838, May 1992.

[53] T. Yang and D. H. K. Tsang, "A Novel Approach to Estimating the Cell Loss Probability in an ATM Multiplexer Loaded with Homogeneous On-Off Sources", IEEE Transactions on Communications, Vol.COM-43, No. 1, pp:117-126, January 1995.

[54] S. H. Kang and D. K. Sung, "A CAC Scheme Based on Real-Time Cell Loss Estimation for ATM Multiplexers", IEEE Transactions on Communications, Vol.COM-48, No. 2, pp:252-258, February 2000.

[55] Box, G. E. P., and Jenkins, G. M., "*Time Series Analysis, Forecasting and Control*", Holden Day, San Franciso, California, 1970.

[56] D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources", Bell System Technical journal, Vol. 61, No. 8, pp: 1871-1894, Oct. 1982.

[57] B. Maglaris et.al., "Performance Models of Statistical Multiplexing in Packet Video Communications", IEEE Transactions on Communications, Vol.COM-36, July 1988, pp. 834-844.

[58] J. Cosmas et. al., "A Review of Voice, Data and Video Traffic Models for ATM", European Transactions on Telecommunication, Special Issue on Teletraffic Research for B-ISDN in the RACE program, Vol. 5, no 2, 1994, pp. 139-154.

[59] P. Sen et.al., "Models for Packet Switching of Variable-Bit-Rate Video Sources", IEEE Journal on Selected Areas in Communications, Vol. 7, No. 5, June 1989, pp:865-869.

[60] Mark W. Garrett, "A Service Architecture for ATM: From Applications to Scheduling", IEEE NEtwork, May/June 1996, pp:6-14.

[61] J. Sanchez et. al., "A Survey of MAC Protocols Proposed for Wireless ATM", IEEE Network, Nov./Dec. 1997, pp. 52-62.

[62] John Porter, "ORL Radio ATM", http://www.cam-orl.co.uk

[63] M. Umehira et.ai., "ATM Wireless Access for Mobile Multimedia: Concept and Architecture", IEEE Personal Communications, Oct. 96, pp: 39-48.

[64] H. Kist and D. Petras, "Service Strategy for VBR Services at an ATM Air Interface", Second European Mobile Communications Conference, Bonn, Germany, Oct. 1997.

[65] D. Petras et. al., "Support of ATM Service Classes in Wireless ATM Networks", ACTS Mobile Communications Summit, Aalborg, Denmark, Oct. 1997.

[66] C. S. Chang et. al., "Guaranteed Quality-of-Service Wireless Access to ATM Networks", IEEE Journal on Selected Areas in Communications, Vol.15, No. 1, Jan. 1997, pp:106-118.

[67] M. Shafi et. al., "Wireless Communications in the Twenty-First Century: A perspective", Proceedings of the IEEE, Vol. 85, No. 10, October 1997, pp: 1622-1638.

[68] Jun Huang, Tho Le-Ngoc and Jeremiah F.Hayes, "A Broadband Satellite Communications System for Multimedia Services", International Conference on Communications, June 23-27, Dallas, Texas, 1996, pp: 906-909.

[69] A. Baiocchi et. al., "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed On-Off Sources", IEEE JSAC, Vol. 9, No. 3, April 1991, pp: 388-393.

[70] Hiroshi Saito, "Departure Process of an N/G/1 Queue", Performance Evaluation 11, 1990, pp: 241-251.

[71] Ricardo Gusella, "Characterising the Variability of Arrival Processes with indexes of Dispersion", IEEE Journal on Selected Areas in Communications, Vol. 9, No. 2, Feb. 1991, pp:203-211

[72] Tatsuya Takine, et. al., "Cell Loss and Output Process Analyses of a Finite-Buffer Discrete-Time ATM Queueing System with Correlated Arrivals". IEEE Infocom'93. San Franciso, California, U.S.A.. March 28 - April 1, 1993, pp: 1259-1269.

[73] Nelson L. S. Fonseca and John A. Silvester. "Modelling the Output Process of an ATM Multiplexer with Markov Modulated Arrivals". IEEE International Conference on Communications'94. New Orleans. Louisiana, U.S.A.. May 1 - 5, 1994, pp:721-725.

[74] Nelson L. S. Fonseca and John A. Silvester. "Modelling the Output Process of an ATM Multiplexer with Correlated Priorities". IEEE International Conference on Communications'97. Montreal. Quebec. Canada, June 8 - 12, 1997, pp:816-821.

[75] P. T. Brady, "A Statistical Analysis of On-Off Patterns in 16 Conversations". Bell Syst. Tech. Journal, Vol. 47, No. 1, Jan. 1968, pp: 73-91.

[76] J. S. Kaufman, "Blocking in a Shared Resource Environment". IEEE Transactions on Communications, Vol.COM-29, oct-1981, pp. 1474-1481.

[77] G. Stamatelos and Jeremiah F.Hayes, "Blocking in Tandem ISDN Links", IEEE Transactions on Communications, Vol. COM-41, August 1993, pp. 1252-1259.

[78] M. F. Neuts, "A Versatile Markovian Point Process". *Journal of Applied Probability*, 16, 1979, pp: 764-771.

[79] M. F. Neuts, "The C-server Queue with Constant Service Times and a Versatile Markovian Arrival Process", *Applied Probability-Computer Science: The interface*, 1, Jan. 5-7, 1981, pp:31-67.

[80] M. F. Neuts, "Moment Formulas for the Markov Renewal Branching Process", Advances in Applied Probability, Vol. 8, 1976, pp: 690-711.

[81] M. F. Neuts, "**Structured Stochastic Matrices of M/G/1 Type and their Applications**", Marcel Dekkar Inc., 1989, Chapter 5, pp: 310-329.

[82] M. F. Neuts, "Queues Solvable without Rouche's Theorem". Operations Research, 1979. pp: 767-781.

[83] C. Blondia. "The N/G/1 Finite Capacity Queue". Commun. Stat.-Stochastic Models, Vol. 5, no. 2, pp: 273-294, 1989.

[84] H. Heffes and D. M. Lucantoni. "A Markov Modulated Characterisation of Packetised Voice and Data Traffic and Related Statistical Multiplexer Performance". IEEE Journal on Selected Areas in Communications. Vol. SAC-4. No. 6, Sep. 1986. pp: 856-868.

[85] Raif O. Onvural. "**Asynchronous Transfer Mode Networks Performance Issues**". Artech House Inc., 1994.

[86] David M. Lucantoni and V. Ramaswami. "Efficient Algorithms for Solving the Non-linear Matrix Equations Arising in Phase type Queues". Stochastic Models. Vol. 1 (1). 1985, pp: 29-51.

[87] David M. Lucantoni, "New Results on the Single Server Queue with a Batch Markovian Arrival Process". Commn. Statist. - Stochastic Models, 7 (1), 1991. pp:1-46.

[88] Dario Bini and B. Meini. "On the Solution of a Non-linear Matrix Equation Arising in Queueing Problems". SIAM Journal on Matrix Analysis and Applications. Vol. 17, No. 4, October 1996, pp: 906-926.

[89] Thimma V.J. Ganesh Babu, Tho Le-Ngoc and Jeremiah F. Hayes, "A Unified Approach for Evaluating Call Blocking and Burst Blocking in High Speed Networks", Canadian Conference on Electrical and Computer Engineering, 5-8 September, Montreal. Quebec, Canada. 1995, pp. 846-849.

[90] Thimma V.J. Ganesh Babu. Tho Le-Ngoc and Jeremiah F. Hayes. "Performance Analysis of On-board Switching in Broadband Satellite Communication Systems". International Conference on Communications. Dallas. Texas. June 23-27. 1996, pp: 1487-1491.

[91] Thimma V.J. Ganesh Babu. Tho Le-Ngoc and Jeremiah F. Hayes. "Performance Evaluation of Priority based Service in Multimedia ATM Networks". 2nd. IFIP Workshop on Traffic Management and Synthesis of ATM networks. Montreal. September 24-26. 1997.

[92] Thimma V.J. Ganesh Babu. Tho Le-Ngoc and Jeremiah F. Hayes. "Performance of A Priority-Based Dynamic Capacity Allocation Scheme for WATM Systems". Globecom. Sydney, Australia. 8-12 November. 1998.

[93] Thimma V.J. Ganesh Babu. T. Le-Ngoc, Jeremiah F. Hayes. "Performance Analysis of OBP based Multimedia Multibeam Satellite Networks". International Conference on Communications. Atlanta, Georgia. June 7-11. 1998. pp: 503-507.

[94] Thimma V.J. Ganesh Babu. Tho Le-Ngoc and Jeremiah F. Hayes. "Performance Evaluation of A Wireless ATM Switch Using Priority Based Dynamic Capacity Allocation Scheme", Globecom. San Francisco, California. November 27 - December 1. 2000.

[95] Thimma V.J. Ganesh Babu. Tho Le-Ngoc and Jeremiah F. Hayes. "Performance of A Priority Based Dynamic Capacity Allocation Scheme for Wireless ATM

Systems", IEEE Journal on Selected Areas in Communications, Vol. 19, No. 2, February 2001, pp: 355-369.

[96] Thimma V.J. Ganesh Babu, Tho Le-Ngoc and Jeremiah F. Hayes, "Performance Evaluation of A Wireless ATM Switch Using Priority Based Dynamic Capacity Allocation Scheme", under review for IEEE Transactions on Communications.

[97] Shahram Shah-Heydari, "MMPP modeling of ATM Multimedia Traffic", M.A.Sc Thesis. Dept. of E.C.E., Concordia University, July 1998.