Traffic Management of the ABR Service Category in ATM Networks

PhD Thesis

Llorenç Cerdà i Alabern

Barcelona, Octubre de 1999

Traffic Management of the ABR Service Category in ATM Networks

Barcelona, Octubre de 1999

Universitat Politècnica de Catalunya
Departament d'Arquitectura de Computadors

Traffic Management of the ABR Service Category in ATM Networks

PhD Thesis
Llorenç Cerdà i Alabern

PhD Advisor:

Prof. Dr. Olga Casals Torres	Universitat Politècnica de Catalunya, Spain

Members of the PhD tribunal:

Dr. José María Barceló Ordinas	Universitat Politècnica de Catalunya, Spain
Prof. Dr. Vicente Casares Giner	Universitat Politècnica de València, Spain
Prof. Dr. Ulf Körner	Lund University, Sweden
Prof. Dr. Ramon Puigjaner Trepat	Universitat de les Illes Balears, Spain
Prof. Dr. Ioannis Stavrakakis	University of Athens, Greece

Contents

Preface and Acknowledgments

The work included in this PhD thesis was initiated in March'94 when Jorge García and Olga Casals came to my office and proposed to me to join their research group. At that time I had started working as lecturer at their department, the Computer Architecture Department at the Polytechnic University of Catalonia. My only background was the studies of Telecommunication Engineer at the same University.

I accepted the challenge of Jorge and Olga with enthusiasm. The research field was the traffic management in ATM networks. Although ATM networks were unknown for me, I had some learning about computer networks from my career and I was very motivated to extend my knowledge on this field.

During my research I have had the opportunity to deal with different evaluation tools: simulation, analytical and experimental with real equipment. I have written my own ATM simulator in C++ to carry out all the simulation results presented in this thesis. The program started with a small event driven program, but it constantly grow during all my research and now is formed by more than 10,000 lines of code. I have spent many hours adding new features and debugging the simulator. Actually, I have experienced the "fever of computers", writing code without feeling the time passing until I got the desired feature working in my program. I have discovered that a lot of creativity and imagination can be experienced in programming. Furthermore, getting the desired results without any *core* being created by operating system is a great pleasure.

Performance evaluation by means of analytical models has been also very stimulating. Stochastic processes and Markov chains are very beautiful fields of mathematics, and require a big deal of deductive reasoning and inventiveness for modeling complex systems as those covered in this PhD. Deriving a successful mathematical model is often more uncertain and arduous than writing a program to obtain the same results by simulation. However, a great gratification is obtained when the analytical model gives the desired outcome.

Finally, I had the opportunity to experiment with real equipment at the testbed of the European EXPERT Project, located in Basel. Experimenting with real equipment is also a fascinating experience. Furthermore, it allows to move from the abstract plane of simulation and mathematics to the real world of cables, connectors, switches, generators, etc. Of course, the challenges which arise when dealing with real equipment at the testbed are very different than those of simulation and mathematics. First, one is constrained to the equipment present at the laboratory, which in our case offered a very limited range of configurations. Then, preparing the experiments needs the knowledge of the configuration of the equipment. This usually requires the reading of complicated manuals. Finally, the traces recorded by the analyzer were long ASCII files of time-stamps and contents of the ATM cells. Therefore, an elaborated manipulation with UNIX tools was still needed to derive understandable results from the traces.

Of course, I wouldn't have been able to do all the work contained in this PhD without the help of many people. I want to thank to my research group colleagues: Olga Casals, Jorge García, José María Barceló and Fernando Cerdán. I have maintained many interesting discussions and comments with them. Olga has been my PhD advisor. I want to thank her for all the valuable guidelines during my research, and for patiently proofreading all my work. Jorge has a natural skill for analytical modeling, thus, I directed to him many questions related with this issue. José María is my office mate, so, he has been the first one listening my ideas and questions. I want to thank also to Benny Van Houdt from University of Antwerp. We collaborate very actively in paper [23], and I learned from him a lot about the matrix geometric technique.

I want to thank all the colleagues I had at the EXPERT testbed in Basel: Egil, Martin, Laurent, Fernando, Kathleen, Sandra, Bart, Nicky and Jordi. We collaborated altogether to obtain the traces shown in chapter 6.

Finally I want to thank all my family for their support. Specially to my mother Mercè and my wife Csilla to whom I tried to explain many times the *strange things* I was studying in my PhD.

Chapter 1

Introduction

1.1 B-ISDN

The *Integrated Services Digital Network* (ISDN) was formalized in 1984 when the CCITT standardization body adopted the “I” series recommendations. CCITT is the former name of the *International Telecommunications Union - Telecommunications Standardization Sector* (ITU-T). ITU-T is an international organization whose activities include the coordination, development, regulation and standardization of telecommunications.

The idea behind the ISDN was to develop a network that provided end-to-end digital connectivity to support a wide range of services, including voice and data transfer. Traditionally, different networks dedicated to each of these services had been used.

The original ISDN was based on the digital telephone network, characterized by the 64 kbit/s channel. However, the ISDN hierarchy derived from the multiplexing of 64 kbit/s channels provided a rigid bit rate scheme. This was not acceptable to bear the foreseeable future broadband services. Therefore, a new network, the *Broadband-ISDN* (B-ISDN) was defined to be supported by a completely different technology: the *Asynchronous Transfer Mode* (ATM).

ATM is a connection-oriented technique based on the fast multiplexing and switching of fixed-size packets (48 octets information field and 5 octets header) called cells. Since ATM was introduced an extensive research and standardization effort have been done. The ITU-T and later the *ATM Forum* have led the standardization of ATM.

The ATM Forum is a private, non-profit industry consortium made up of manufacturers, carriers and end user whose purpose is to develop and outline standards for ATM.

1.2 Dealing with the Integration of Services

In order to fulfill the B-ISDN requisites, the ATM technology has to provide the transport facilities to efficiently support a diversity of traffic classes (e.g. voice, video, and data). Each traffic class may have different *Quality of Service* (QoS) requirements. Moreover, applications may have very different traffic characteristics.

To cope with this diversity of requirements the concept of *Service Categories* (SCs) has been defined by the ATM Forum. In the ITU-T terminology these are referred as *ATM Transfer Capabilities* (ATCs).

At the connection setup the source has to chose one of the Service Categories implemented by the network. Each SC may have a different QoS and traffic parameter set to be negotiated. The source should then submit the traffic according to the rules specified for the chosen SC.

1.3 Topics Addressed in this Thesis

When I started my PhD research I had in mind the study of the ATM traffic control schemes designed for data traffic to be considered for SC/ATC standardization. At that time both ITU-T and the ATM Forum were looking for an efficient scheme for this kind of traffic.

My first work was a performance evaluation study of the *Fast Reservation Protocol* (FRP). FRP was proposed by some France Telecom engineers [9]. The basic idea behind the FRP protocol is a kind of connection acceptance control at burst level. That is, when a source wants to transmit a burst it is accepted or blocked depending on the available bandwidth within the link. When a burst is blocked successive re-attempts are made until it is accepted. Later on, the FRP principles were used by the ITU-T to standardize the *ATM Block Transfer* (ABT) ATC.

Meanwhile, industries grouped into the ATM Forum were considering two different possibilities for the *Available Bit Rate* (ABR) SC [62]: a credit-based [49, 50] and a rate-based [80]. Both schemes have in common that sources dynamically regulate the cell rate using feedback information from the network. The credit-base scheme is based on a link-by-link window flow control mechanism. Independent flow controls are performed on each link, and each connection must obtain buffer reservation for its cell transmission on each link. A connection is allowed to continue the cell transmission as long as it gains credits from the next node. This feedback mechanism allows transient congestion to be relieved effectively and avoids cell losses.

On the other hand, the rate-based scheme consists of the switches directly adjusting the cell rate of the contending sources to the available bandwidth of the links. In order to do that the sources transmit the so called *Resource Management* cells (RM-cells) which are *turned around* along the same path by the destination. These RM-cells are used by the switches to convey the rate changes to the sources.

After strong discussions between the ATM Forum members, the rate based approach was accepted to support ABR. The main arguments for that decision where: (i) the complex buffer allocation that a credit-based approach would require in a wide area network, (ii) the rate-based scheme permits a variety of switch implementations. This makes possible a diversity of products with different degrees of complexity and performance, maintaining compatibility with the standard.

The ATM Forum specification of ABR appeared first in April 1996 in the “ATM Forum Traffic Management Specification Version 4.0” [4]. I dedicated the major part of my PhD to study different issues related with the ABR using the framework described in this specification.

At the moment of finishing my PhD, ABT and ABR are not the only ATM traffic control schemes specified for an efficient data traffic transport. In the “ATM Forum Traffic Management Specification Version 4.1” [5] approved by the ATM Forum in March 1999 the *Guaranteed Frame Rate* (GFR) Service Category has been added to the previous specification. The origins of GFR go back to December 1996 when the Service Category proposal called *UBR+* was made in the ATM Forum [34]. This proposal was mainly motivated because a major part of the current data traffic, e.g. the Internet traffic, is generated by users that may not have equipment able to comply with the source specification required for ABR. These users would be forced to use a SC not appropriated by their applications, and therefore, they would have little or no incentive to migrate to ATM. GFR has been conceived to offer an easy access to ATM for these kind of users.

However, the study of the GFR Service Category goes beyond the scope of this PhD. As mentioned before, I devoted the larger part of my PhD to ABR, covering the following topics:

- switching mechanisms,
- conformance definition,
- charging,
- ABR support to TCP traffic.

I studied switching mechanisms from different points of view. In my first work on ABR I analyzed several switch algorithm proposals. To do so I developed an event driven simulator written in C++. I also looked for improvements related to switch algorithms. Finally, I had the opportunity do derive some conclusions from experiments developed with real equipment at the EXPERT Testbed [27].

The conformance definition is the formalism used by the network to monitor whether the sources transmit according to the traffic contract. The conformance algorithm standardized for ABR is the *Dynamic Generic Cell Rate Algorithm* (DGCRA) . In my first work about the DGCRA I analyzed by simulation the problems related with the algorithm and I proposed some improvements. I also investigated the DGCRA parameter dimensioning using analytical and simulation techniques.

Another topic I investigated is the charging of ABR. The charging scheme that will be applied has not been specified. Pricing however may be an essential condition for the users when submitting traffic. I analyzed existing schemes and I proposed new alternatives.

I concluded my PhD research analyzing the ABR support to TCP traffic. This research is motivated because the Internet traffic using the TCP/IP set of protocols is currently the biggest part of non real time traffic. Therefore, ABR may be one of the Service Categories chosen to give support to the Internet traffic over ATM networks.

1.4 Outline

In the following I outline the contents of each of the remaining chapters. I also give the references to the related reports and papers I contributed while developing the thesis.

Chapter 2 "Fundamentals of ATM" explains the ATM background needed to understand this thesis. This includes an overview of B-ISDN and traffic management in ATM Networks. The Service Categories specification and the network functions designed to guarantee the QoS are described.

Chapter 3 "Performance Evaluation of the ATM Block Transfer (ABT)" focuses on the analysis of the ABT fairness when different sources are multiplexed. The work included in this chapter was presented in [22]. Together with a selection of the papers presented in this conference, the paper was accepted to be published in [48].

The analytical analysis assumes an ON-OFF model for the data sources with exponential ON and OFF time distribution (burst-silence model). Two approximations of the protocol are considered that allow a Markov chain model to be considered. The first approach gives an approximation of the burst blocking probability and blocking time. However, the Markov chain does not have a product form solution and the stationary probabilities have to be obtained numerically solving the global balance equations. In the second approach further approximations lead to a Markov chain which has a product form solution. This allows us to approximate the burst blocking probability by a simple formula. Analytical results are validated by simulation.

Chapter 4 "The Available Bit Rate (ABR) Service" gives a detailed description of the ABR-SC. This is the theoretical background needed to understand the following chapters.

Chapter 5 "Switching Mechanisms for ABR" performs a comparison of ABR switch algorithms by simulation. This chapter includes the first ABR work I did in [12] and my contribution appeared in the EXPERT deliverable [28].

Chapter 6 "Experimental Analysis of an ER Switch" analyzes an ABR switch performance by means of a set of experiments carried out at the expert Testbed. This work was published in [18].

Chapter 7 "Improvements of ER Switch Algorithms" describes the paper presented in [15]. This paper shows that ABR switch algorithms could be improved at the cost of some minor modifications of the ATM Forum standard. These results are obtained by simulation.

Chapter 8 "DGCRA: The Conformance Definition for ABR" gives a detailed description and analysis of the ABR conformance definition (the DGCRA) and will include the related work published in [13] and [10].

The DGCRA analysis presented in these publications is performed by simulation. It shows the problems that the conformance definition may have. Improvements are also proposed to solve the addressed problems.

Chapter 9 "DGCRA Parameter Dimensioning" includes the results presented in [16, 19, 23]. In these contributions the DGCRA is analyzed applying analytical and simulation techniques.

In [16, 19] a discrete time approach which leads to an iterative algorithm is used. This approach was previously used to analyze the conformance definition for the CBR-SC [65]. It is a semi-analytical method based on a renewal assumption. The input traffic is characterized by a general independent distribution that can be obtained e.g. by simulation.

In [23] we use another analytical approach based on a matrix geometric technique. In this case the framework is simpler, however, a good analytical characterization of the DGCRA is obtained.

Chapter 10 "Charging of ABR" studies the charging of ABR. It includes the contributions [21, 14, 17, 29]. In these contributions charging schemes have been classified as "static" and "dynamic". Static refers to those pricing schemes where the charging parameters are established at the connection set up and do not change afterwards. Dynamic, instead, is used when the charging scheme changes the prices according to source demands.

We first analyze a static scheme proposed by Songhurst and Kelly [73]. The drawbacks of the scheme are discussed and a new alternative which tries to solve them is proposed. Then, a dynamic scheme proposed by Courcoubetis et al. [25] is considered. Again, the drawbacks are discussed and a new approach is presented. Moreover, this new approach is evaluated by means of an analytical fluid model. Finally, a numerical comparison obtained by simulation shows the pros and cons of the different charging schemes.

Chapter 11 "ABR Support to TCP" analyzes by means of simulations the iteration between the TCP protocol used in Internet and ABR. The simulations consider the interconnection of LANs with different scenarios: using shared media LANs and ATM-LANs. ABR is confronted with the simpler UBR service category by giving averages and plots of the traces of the most representative parameters. These include goodput, buffer occupancy, allowed window of the TCP module etc. The simulation results give guidelines about the TCP and ABR/UBR inter-operation and the benefits of using ABR over UBR. This work is included in [20].

Chapter 2

Fundamentals of ATM

2.1 Introduction

This chapter surveys the ATM background needed to understand this thesis. The intention of this chapter is to highlight the main topics in order to be a point of reference for the rest of the thesis.

The chapter is organized as follows. In section 2.2 the basis of the ATM technology are introduced. Remember from section 1.1 that ATM was born to give support to B-ISDN. Both concepts are intimately related (and often used indistinctively). Section 2.3 overviews the standardized access interface to the B-ISDN. Section 2.4 introduces the concept of *Traffic Management* and related topics like the *Traffic Contract* and the *Quality of Service* (QoS) parameters. Finally, section 2.5 describes the *Service Categories* that has been defined by the ATM Forum.

2.2 The Asynchronous Transfer Mode (ATM)

Optical fiber facilities, which were deployed in public network inter-office trunks in the early 1980s are expected to penetrate access distribution networks. The ATM technology has been conceived to operate in these fiber–based networks running at high rates. E.g. two possible bit rates in the order of 155 Mbps and 622 Mbps have been defined for the *User Network Interface* (see section 2.3).

In ATM all information to be transferred is allocated into fixed-size packets called cells. These cells have a 48 octet information field and 5 octet header. The use of these short cells and the high bit rates result in transfer delays and delay variations which are sufficiently small to enable it to be applied to real–time services such as voice and video.

Furthermore, ATM is a connection-oriented low overhead concept of virtual channels which has no flow control or error recovery. ATM guarantees the cell sequence integrity of cells belonging to a specific virtual channel. Two hierarchical levels are defined:

- *Virtual Channel* (VC): used to describe unidirectional transport of ATM cells associated by a common unique identifier value. This identifier is called the *Virtual Channel Identifier* (VCI).

- *Virtual Path* (VP): used to describe unidirectional transport of cells belonging to virtual channels that are associated by a common identifier value. This identifier is called the *Virtual Path Identifier* (VPI).

2.3 B-ISDN User Network Interface

The *User Network Interface* (UNI) has been defined in the B-ISDN (see figure 2.1). This interface standardizes the access procedures from the *Customer Premises Network* (CPN) to the public ATM Network. The CPN is the network located at the user side of the B-ISDN.

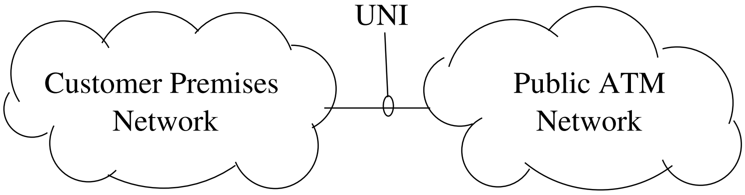


Figure 2.1: B-ISDN Interfaces.

The set of protocols which define the UNI are organized with the reference model shown in figure 2.2.

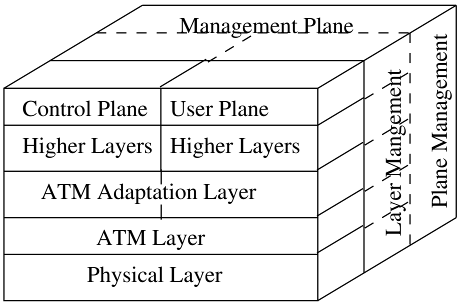


Figure 2.2: B-ISDN Protocol Reference Model.

This reference model consists of three planes:

- *User Plane*: provides for the transfer of user information.
- *Control Plane*: is responsible for the call control and connection control functions. These are all signaling functions which are necessary to set up, supervise and release a call or connection.
- *Management Plane*: includes two types of functions called *Layer Management* and *Plane Management* functions. The Layer Management handles the specific *Operation And Maintenance* (OAM) information flows for each layer. OAM actions are responsible for the monitoring and supervision of the network. Examples are: performance monitoring of network entities, failure detection, fault localization, etc.

The management functions that relate to the whole system are located in the Plane Management. This is responsible for providing coordination between all planes. No layered structure is used within this plane.

In the following the layer functionalities defined within the planes are described.

2.3.1 ATM Adaptation Layer (AAL)

This layer adapts the service which is provided by the ATM layer to a service that can better support specific classes of applications. One of its functions is the *Segmentation and Reassembly*. This consists of the segmentation of the data of the higher layer into ATM cells, and the reassembly of the payload of ATM cells into a format the higher layer can understand.

In order to specify the AAL protocols the following classes of services were defined by the ITU-T:

	Class A	Class B	Class C	Class D
Timing Requirements	Required		Not Required	
Bit rate	Constant	Variable		
Connection Mode	Connection Oriented			Connectionless

Table 2.1: Service classification for the ALL.

Four types of AALs have been defined to support these classes of services. In the following, examples of services of each of these classes and the corresponding AAL are given:

- Class A examples include 64 Kbps voice, fixed-rate uncompressed video and leased lines for private data networks. AAL1 has been designed to accommodate this kind of services. E.g. handling of cell delay and delay variation is performed by AAL1.
- Class B examples include compressed voice or video. The AAL2 is devoted to these kind of services.
- Class C examples include data network applications where a connection is set up before data is transferred, e.g. file transfer. The ITU-T originally recommended two types of AAL to support these services, but they have been merged into a single type called AAL3/4. Another AAL, called AAL5, has also been introduced. AAL5 offers similar functionalities to AAL3/4, but simplifying the protocol procedures.
- Class D includes data network applications where no connection set up is needed before data is transferred. Either AAL3/4 or AAL5 are suitable for these kind of services.

Although each AAL has been designed for a specific type of service class, the standard allows a free mapping of service classes and AAL types.

2.3.2 ATM Layer

The functionalities of this layer are given by the ATM cell header (see figure 2.3).

Functionalities related to VCI/VPI are the multiplexing and demultiplexing and VCI/VPI translations at the switches (see section 2.2). The other fields and related purposes are described in the following:

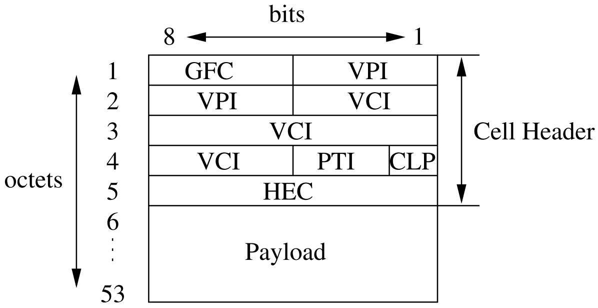


Figure 2.3: ATM cell format.

PTI	Cell type
000	User data cell, EFCI = 0, AUU = 0
001	User data cell, EFCI = 0, AUU = 1
010	User data cell, EFCI = 1, AUU = 0
011	User data cell, EFCI = 1, AUU = 1
100	OAM cell
101	OAM cell
110	Resource Management cell (RM-cell)
111	Reserved

Table 2.2: PTI values.

- *Generic Flow Control* (GFC) has been designed as a means of controlling the traffic flow from ATM connections at the UNI. E.g. it can be used to control the access flow to a shared link.

- *Payload Type Indicator* (PTI) is a three bits field which identifies the type of information carried in the ATM cell. Table 2.2 shows the possible values for the PTI field. In the following the identifiers shown in the table are described.

  – EFCI stands for *Explicit Forward Congestion Indication* and maybe used to indicate congestion by some traffic control mechanisms (see chapter 4).

  – AUU stands for *ATM layer user to ATM layer user*. This indicator is used by AAL5 to delimit the AAL5 protocol data units within the ATM cell stream. This is done by transmitting all the ATM cells as AUU = 0 except the last cell belonging to an AAL5 protocol data unit, which is transmitted with AUU = 1.

  – OAM indicates an Operation And Maintenance cell (see section 2.3).

  – RM-cells are used by some traffic control mechanisms as the ABR (see chapter 4) and ABT (see chapter 3).

- *Cell Loss Priority* (CLP) is a single bit field which serves a dual purpose:

  1. Setting of the CLP (CLP=1) by the sending terminal to signify that the cell carries nonessential information (and thus that the cell is selectively discardable under congestion conditions).

2. Setting the CLP during access to the network if the *Usage Parameter Control* (UPC) finds that the cell is not in conforming with the traffic contract (see section 2.4). The standard specifies that a non conforming cell can be either marked (by setting CLP=1) or discarded by the UPC.

- *Header Error Check* (HEC) field contains the result of an 8-bit CRC checking on the ATM header (but not on the data). A single bit can be corrected with this code.

2.3.3 Physical Layer

This layer is responsible for the bit timing and other functions for the suitable transmission of the electrical/optical signals in the transmission media. In addition, it provides cell delineation and HEC generation and processing.

2.4 Traffic Management

The primary role of traffic management is to protect the network and end systems from congestion so that the *Quality of Service* (QoS) commitments can be maintained. An additional role is to promote the efficient use of network resources. ATM technology is intended to support a variety of applications with a diversity of QoS requirements. To meet these objectives, a set of *Generic Functions* for managing and controlling traffic are defined. Among others, the following Generic Functions have been specified:

- *Connection Admission Control* (CAC): Is performed at the call set-up. This function checks whether the available network resources can accommodate a new connection request. The CAC may accept or reject a new connection.

- *Usage Parameter Control* (UPC): Is performed all along the duration of the connection. It monitors and control the cell stream of the connection such that the parameters negotiated at the connection set-up are respected.

Furthermore, ATM standardization bodies have defined specific categories tailored for different traffic and QoS requirements. These are referred to as *Service Categories* (SCs) by the ATM Forum and *ATM Transfer Capabilities* (ATCs) by the ITU-T.

It is mandatory that the SC/ATC used on a given ATM connection be implicitly or explicitly declared at connection set–up. Each SC/ATC may have a different QoS and traffic parameter set to be negotiated. The source should then submit the traffic according to the control functions specified for the chosen SC. The set of parameters and agreements negotiated at the connection set up is called the *Traffic contract*.

In the thesis I will use the terminology formulated by the ATM Forum. The reason for this choice is that the major topic of this thesis, the ABR SC, has been mainly developed by the ATM Forum and is defined only partly in the ITU-T documents. I will only use the ITU-T terminology when I'll refer to the *ATM Block Transfer* ATC. This is the only ATC that cannot be mapped into a corresponding SC. Table 2.3 shows the correspondence between the ITU-T ATCs [35] and the ATM Forum SCs [5].

ITU-T ATCs	ATM Forum SCs
Deterministic Bit Rate (DBR)	Constant Bit Rate (CBR)
Statistical Bit Rate (SBR)	Variable Bit Rate (VBR)
not specified	Unspecified Bit Rate (UBR)
Available Bit Rate (ABR)	Available Bit Rate (ABR)
not specified	Guaranteed Frame Rate (GFR)
ATM Block Transfer (ABT)	*not specified*

Table 2.3: Correspondence between the ITU-T ATCs and the ATM Forum SCs.

2.4.1 Traffic Contract

During the connection setup procedure the end system chooses a SC and a QoS is negotiated. Furthermore, a set of parameters is needed by the network in order to estimate the required resources to be allocated, or reject the request if they are not available. These parameters are referred to as the *Connection Traffic Descriptor.* Acceptance of the connection obliges the network to provide the specified QoS and throughput. Additionally, the user is obliged to limit its traffic according to the given parameters. This agreement forms the basis of the so–called *Traffic Contract* between the network and the user.

Therefore, the Traffic Contract parameters are mainly used by the CAC function to allocate resources and select routes, and by the UPC to decide is the connection cell stream is conforming. To meet these objectives, the Connection Traffic Descriptor is composed of a *Source Traffic Descriptor* and the additional information for the conformance testing of cells by the UPC. The Source Traffic Descriptor is defined as the set of parameters of the ATM source that allows to be quantitatively used as a basis for resource allocation. Figure 2.4 provides the overall view of how the parameters are grouped in the Traffic Contract. These parameters are described in the following.

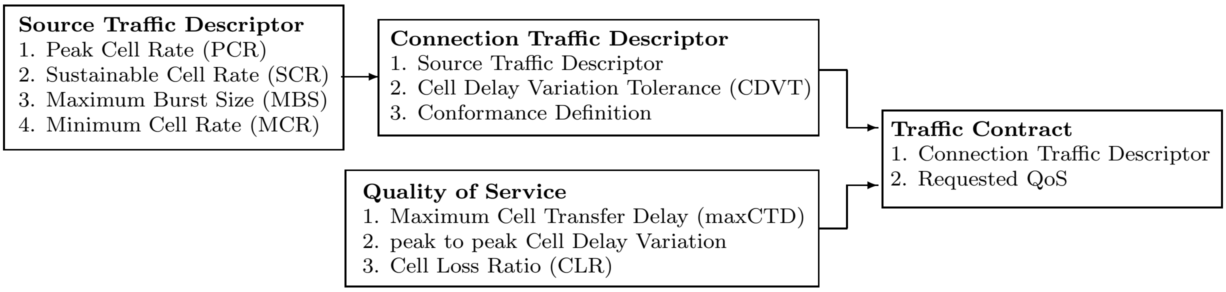

Source Traffic Descriptor
1. Peak Cell Rate (PCR)
2. Sustainable Cell Rate (SCR)
3. Maximum Burst Size (MBS)
4. Minimum Cell Rate (MCR)
Connection Traffic Descriptor
1. Source Traffic Descriptor
2. Cell Delay Variation Tolerance (CDVT)
3. Conformance Definition
Traffic Contract
1. Connection Traffic Descriptor
2. Requested QoS
Quality of Service
1. Maximum Cell Transfer Delay (maxCTD)
2. peak to peak Cell Delay Variation
3. Cell Loss Ratio (CLR)

Figure 2.4: Traffic parameters sets.

2.4.2 Source Traffic Descriptor

This set of parameters varies depending on the SC chosen by the source (see table 2.4). In the following the list of parameters that may be specified is given, with an informal description of their meaning. Next sections provide a more accurate description.

- *Peak Cell Rate* (PCR) is the maximum instantaneous rate at which the user may schedule cells for transmission.
- *Sustainable Cell Rate* (SCR) is the average rate measured over a long time interval.
- *Maximum Burst Size* (MBS) is the maximum number of back-to-back cells that can be sent at the peak cell rate.
- *Minimum Cell Rate* (MCR) is the minimum rate guaranteed for the ABR SC (see section 2.5.4) and the GFR SC (see section 2.5.5).
- *Maximum Frame Size* (MFS) is the maximum frame size for the frame conformance in the GFR SC (see section 2.5.5).

Figure 2.5 gives an illustrative example of a periodic cell stream with (PCR, SCR, MBS) parameters.

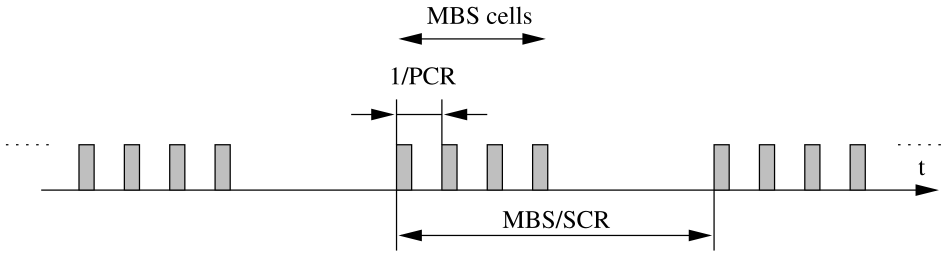


Figure 2.5: Periodic cell stream with (PCR, SCR, MBS) parameters.

Due to the unavoidable variable delays introduced by the ATM layers and multiplexing stages, a tolerance is needed for the above parameters. Note that these variable delays may cause a variation on the time spacing between consecutive cells. Therefore, a measuring point located at the UNI may see a higher instantaneous rate than the transmission rate used by the end system.

This tolerance leads to an operational definition of the above parameters in terms of the *Generic Cell Rate Algorithm* (GCRA) which is described in the following.

2.4.3 The Generic Cell Rate Algorithm (GCRA)

The GCRA is defined with two parameters: the Increment (I) and the Limit (L). The notation GCRA(I, L) is used to specify this algorithm with increment I and limit L. The algorithm is intended to check the conformance of a cell stream arriving at a measuring point. The cell stream is assumed to have an intercell spacing of I time units on the average, and a tolerance of L time units.

The GCRA checks the cell conformance computing the so called *theoretical arriving time* c_k at each cell k arrival epoch a_k. The interpretation of c_k is the time at which cell k should have arrived according to GCRA(I, L). Therefore, the value $y_k = c_k - a_k$ is the CDV of cell k with respect to c_k. E.g. if $y_k < 0$ the cell k arrives “later” than expected, and arrives “earlier” if $y_k > 0$. Notice that in case of an “earlier” arriving cell, the instantaneous rate seen at the measuring point is higher than expected. The limit L is used as the tolerance for this mismatch, and thus, the cell is considered *conforming* if $y_k \leq L$ and non conforming if $y_k > L$.

The computation of the theoretical arrival time c_k is given by the following algorithm. Let $\{a_k\}_{k\geq 0}$ be the set of arrival times of the cell sequence $\{k\}_{k\geq 0}$. With the initialization $c_0 = a_0$, c_k is defined recursively as:

$$\begin{aligned} y_k &= c_k - a_k \\ c_{k+1} &= \begin{cases} \max(c_k, a_k) + I & \text{if} \quad y_k \leq L, \quad \text{(cell conforming)} \\ c_k & \text{if} \quad y_k > L, \quad \text{(cell non conforming)} \end{cases} \end{aligned} \tag{2.1}$$

2.4.4 Operational Definition of the PCR/SCR

In order to specify the PCR taking into account the unavoidable CDV from the emitting end system to the UNI, the reference model of figure 2.6 has been defined. The model is intended to be general enough to include any implementation of the end system.

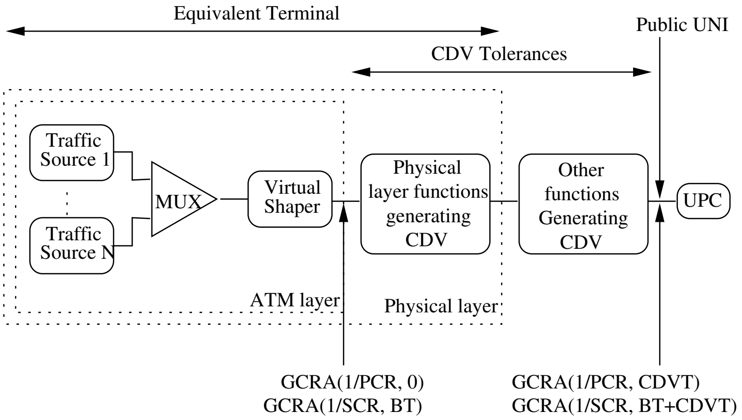


Figure 2.6: PCR/SCR Reference Model.

The virtual shaper represents the imaginary point where the conformance with GCRA(1/PCR, 0) could be applied for the PCR (i.e. where the minimum space between consecutive cells would be 1/PCR time units). The equivalent terminal shown in the figure represents the end system. This is assumed to add a certain CDV at the cell stream. Finally, the overall CDV up to the UNI is represented by the *CDV Tolerance* (CDVT). The CDVT is a mandatory parameter to be specified for the definition of the PCR. Given the PCR and the CDVT, a cell is said to be conforming if test is conforming when the GCRA(1/PCR, CDVT) is applied.

The GCRA is also used to give an operational definition of a variable cell rate stream by means of the PCR, SCR and MBT parameters. A cell stream is considered compliant with these parameters when its cells are conforming when the GCRA(1/SCR, BT) is applied. BT stands for *Burst Tolerance* (in the ITU-T terminology it is referred to as *Intrinsic Burst Tolerance*, IBT). The relation of BT with the other parameters is given by:

$$\text{BT} = \lceil (\text{MBS} - 1)\ (1/\text{SCR} - 1/\text{PCR}) \rceil \tag{2.2}$$

$$\text{MBS} = 1 + \left\lfloor \frac{\text{BT}}{1/\text{SCR} - 1/\text{PCR}} \right\rfloor \tag{2.3}$$

The previous formulas are obtained by deriving the limit BT required for a periodic cell stream as the one showed in figure 2.5 to be conforming with GCRA(1/SCR, BT).

In order to take into account the CDV introduced into the variable cell rate stream, the CDVT is also a mandatory parameter to be given together with the PCR, SCR and MBT. Then, GCRA(1/SCR, BT+CDVT) is also used for cell conformance.

For a more detailed discussion of the CDV dimensioning the reader can consult e.g. [66, 33].

2.4.5 Connection Traffic Descriptor

The connection traffic descriptor specifies the traffic characteristics of the ATM connection. As shown in figure 2.4, the connection traffic descriptor includes the source traffic descriptor, the CDVT (explained in sections 2.4.2 and 2.4.3 respectively), and the *conformance definition.*

The conformance definition may be defined as the formalism used by the network to monitor whether the source transmits according to the traffic contract. For example, the GCRA described in section 2.4.3 has been standardized by the ATM Forum as the conformance definition for the CBR-SC. Chapters 8 and 9 give a detailed description and analysis of the conformance definition defined for ABR.

2.4.6 Quality of Service Parameters

QoS parameters are probabilistic in nature and may vary over the duration of the connection. The ATM Forum establishes that QoS commitments expected from the network can be evaluated over the long term and over multiple connections with similar QoS commitments.

In order to specify the delay related QoS parameters, the *Cell Transfer Delay* (CTD) measurement has been defined. This is the elapsed time between a cell is transmitted by the source until it arrives at the destination. Figure 2.7 illustrates a typical probability density plot of the CTD of a connection. The figure shows the two delays components of the CTD:

- A *fixed delay* which includes the propagation through the physical media, delays induced by transmission system and fixed components of switch processing.
- A *Cell Delay Variation* (CDV) due to buffering and cell scheduling.

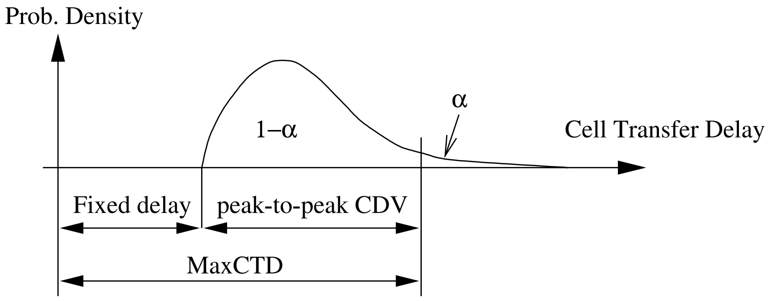


Figure 2.7: Cell Transfer Delay probability density model.

The ATM Forum Identifies three QoS parameters that can be negotiated at the connection set-up (table 2.4 shows which QoS parameters can be negotiated for each SC):

- Peak-to-peak Cell Delay Variation (Peak-to-peak CDV) is the $1-\alpha$ quantile of the CTD minus the fixed CTD (see figure 2.7).

- Maximum Cell Transfer Delay (maxCTD) is the $1-\alpha$ quantile of the CTD (see figure 2.7).

- Cell Loss Ratio (CLR) is defined as the ratio of the lost cells to the total transmitted cells.

2.5 ATM Service Categories

In this section the Service Categories (SCs) introduced in section 2.4 are described. Each of these SCs has been designed to support applications with distinctive traffic characteristics and QoS requirements. Figure 2.8 gives a possible mapping between some typical applications and the corresponding SC. This correspondence is given only for illustrative purposes since the ultimate choice of the user may depend on many issues as: availability of the SC (this may depend on the equipment installed at the source side, and the SCs implemented by the network operator), required QoS, tariffs, etc. A further discussion about the choice of a SC for different application requirements can be found in [67].

Different QoS parameters are allowed to be negotiated for each SC. Furthermore, connection traffic descriptor and traffic management mechanisms differ for each SC. Table 2.4 gives a list of SC attributes (traffic parameters, QoS parameters, and feedback characteristics) that are applied for each SC.

Finally, different *Conformance Definitions* are given for each SC. The Conformance Definition is defined as the formalism used to unambiguously specify the conforming cells of a connection at the UNI. The actual algorithm applied by the UPC is implementation specific. However, QoS commitments of connections compliant with the standardized conformance definition have to be achieved.

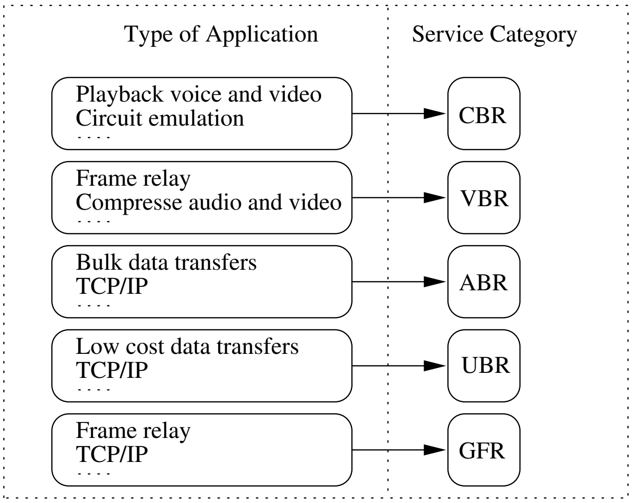


Figure 2.8: Possible mapping between applications and Service Categories.

	ATM Service Category					
	CBR	rt-VBR	nrt-VBR	UBR	ABR	GFR
Traffic Parameters						
PCR	Specified					
CDVT	Specified					
SCR	Unspecified	Specified		Unspecified		
MBS	Unspecified	Specified		Unspecified		Specified
MCR	Unspecified				Specified	
MFS	Unspecified					Specified
QoS Parameters						
peak-to-peak CDV	Specified		Unspecified			
maxCTD	Specified		Unspecified			
CLR	Specified			Unspecified	See Note 1	
Other Attributes						
Feedback	Unspecified				Specified	Unspecified

Note 1: CLR is low for conforming sources.

Table 2.4: ATM Service Category Attributes.

2.5.1 The CBR Service Category

The Constant Bit Rate (CBR) SC roughly provides the service given by a circuit switched network. CBR is intended to support real-time applications requiring tightly constrained cell delay, cell delay variation and cell loss ratio. It is used by connections demanding a static bit rate to be available during the duration of the connection. As shown in table 2.4, only the PCR and the CDVT traffic parameters are requested for this SC, and the conformance definition is given by GCRA(1/PCR, CDVT) (see section 2.4.3).

2.5.2 The VBR Service Category

The Variable Bit Rate (VBR) SC is intended for those applications having a bursty traffic, i.e. having a transmission rate which fluctuates between different rate transmission levels. If the CBR SC was used for this kind of traffic, the PCR should allocate the maximum required transmission rate, and the network resources would be used inefficiently when the source was transmitting at lower rates.

In the VBR SC the sources characterize their traffic such that the network operator do not need to reserve the bandwidth for the maximum transmission rate, but still guaranteeing a certain CLR. The traffic parameters required for the VBR SC are the PCR, SCR and MBS (see section 2.4.2).

The ATM Forum distinguish two VBR sub-categories: real-time VBR (rt-VBR) and non-real-time VBR (nrt-VBR). The rt-VBR is intended for real-time applications requiring tightly constrained delay and delay variation, as voice and video. The nrt-VBR is designed for applications having no delay constrains. As shown in table 2.4, delay related QoS parameters (peak-to-peak CDV and maxCTD) are specified in rt-VBR and unspecified in nrt-VBR.

2.5.3 The UBR Service Category

The Unspecified Bit Rate (UBR) SC is intended for non-real time traffic that do not have delays nor bandwidth constrains. The goal of UBR is to offer an economical SC which employs the unused network resources. As shown in table 2.4, the only required traffic parameter for UBR is the PCR. The network however does not guarantee any QoS.

The applications using this SC are expected to adapt their transmission rate to the time varying network resources in a end-to-end basis.

This SC may seem adequate to give support to other packet networks having end-to-end congestion control as TCP/IP. However, it has been seen that poor performance can be achieved in this kind of scenario (see e.g. [69]). The primary reason is motivated by the fragmentation of the packets in multiple ATM cells. Therefore, the cells dropped in congested switches may produce the useless transmission of many “corrupted packets”.

2.5.4 The ABR Service Category

The Available Bit Rate Service Category (ABR) has been introduced to support traffic from sources which are able to adapt their cell rate to changing network conditions and available bandwidth left by the guaranteed rate traffic. Information about the cell rate adjustments is sent to the sources as feedback information through special control cells, called Resource Management cells (RM-cells).

At the connection setup the source negotiates an upper and lower bounds for the cell rate adjustments. These are respectively the PCR and MCR traffic parameters (see table 2.4). Therefore, the Minimum Cell Rate (MCR) is a guaranteed rate under which the ABR source will be not asked to reduce its transmission rate. The ABR source may negotiate a MCR equal to zero.

For those sources compliant to the ABR feedback control, the network commits a low cell loss ratio and a fair share of the available bandwidth. There is no guarantee with respect to the delay or delay variation. A detailed description of this SC is given in chapter 4.

2.5.5 The GFR Service Category

As mentioned in section 1.3, the Guaranteed Frame Rate (GFR) SC has been mainly conceived to offer an easy migration of the users of the current packet networks, e.g. the Internet, to ATM. This is because these users may not have equipment able to comply with the source specification required for ABR. These users would be forced to use a SC not appropriated for their applications, and therefore, they would have little or no incentive to migrate to ATM.

Unlike the others SCs, the concept of *frame* is introduced in GFR. The term frame refers to a consecutive group of cells conveying a single data unit (e.g. a packet) that has been fragmented at a source end system to be accommodated into the ATM cell stream. Frames are introduced because in case of congestion the network will discard cells in a frame basis in order to avoid delivering “corrupted” packets.

Remember from section 2.3.1 that the AAL is responsible for the *Segmentation and Reassembly* of packets. In order to identify the packets boundaries within the ATM cell stream, the GFR assumes that the AUU indicator of the PTI is used (see section 2.3.2).

In contrast to ABR, no flow control has been defined for GFR. The basic idea of GFR is that a user sending frames that do not exceed the Maximum Frame Size (MFS) in a burst that does not exceed the Maximum Burst Size (MBS), should see the cells delivered across the network with a low cell loss probability.

Additionally to the MBS and MFS, the traffic parameters PCR and MCR are specified in GFR (see table 2.4). The MCR is a minimum cell rate guarantee, defined on a time average basis, to the cell flow compliant with MBS and MFS. Furthermore, the source may transmit at a higher rate up to the PCR. For this excess traffic, the network commits to divide the available resources among the contending sources.

2.5.6 Future ATM Transfer Capabilities

As mentioned before, the current version of the ITU-T I.371 recommendation [35] specifies the DBR, SBR, ABR and ABT ATM Transfer Capabilities (ATCs). DBR, SBR and ABR ATCs are equivalent to the ATM Forum CBR, VBR and ABR Service Categories (see table 2.3), but no equivalent ATCs for UBR and GFR have been defined by the ITU-T.

It is expected than in a new version of the ITU-T I.371 recommendation to appear in March 2000 [36] two new ATCs will be specified, named: Guaranteed Frame Rate (GFR) and Controlled Transfer (CT). The GFR ATC will be equivalent to the GFR SC defined by the ATM Forum explained in the previous section. The CT ATC is a flow controlled by means of credits. As mentioned in section 1.3, the ATM Forum initially considered this possibility for the ABR SC, but it was finally discarded in favor of a rate based flow control scheme.

Chapter 3

Performance Evaluation of the ATM Block Transfer (ABT)

3.1 Introduction

In order to efficiently multiplex data transfers and LAN-LAN interconnection on the ATM B-ISDN, an in-call bandwidth negotiation called Fast Reservation Protocol (FRP) was proposed by France Telecom [9]. Later on, the FRP principles were used by the ITU-T to standardize the *ATM Block Transfer* (ABT) ATC .

The ABT protocol is a kind of Connection Acceptance Control at burst level, that is, when a source wants to transmit a burst it is accepted or blocked depending on the available bandwidth within the link. When a burst is blocked successive reattempts are made until it is accepted.

The performance of an ABT connection is therefore measured in terms of its Burst Blocking Probability (BP) and its Blocking Time (BT, i.e. the time that a blocked burst has to wait until it is eventually accepted). Performance studies of ABT have been carried out by several authors [9], [26], [76], [7]. In those studies however, a set of identical sources is used to model the protocol behavior. When sources with different parameters (PCR and/or burst duration) are multiplexed together, it is foreseeable that each source type will get a different BP and BT. This chapter focuses on the analysis of the ABT fairness when different sources are multiplexed. The term fairness is used in the sense of discrepancy between BP and BT values of different source types. Being all equal, the network would have a fair burst access.

An ON-OFF model is used for the data sources with exponential ON and OFF time distribution (burst-silence model). In order to assess the burst blocking probability of the sources, two approximations of the protocol are considered. In the first approach we assume that the time between reattempts is zero. With different types of sources this case leads to a Markov chain that does not have a product form solution, so it has been analyzed the simple situation in which a set of identical sources are multiplexed with another source of a higher rate.

In a second approach it has been considered that the reattempt time and OFF time are identically distributed. This assumption leads to a Markov chain with a simple product form solution even when considering different source types.

In the first approach the time that a burst has to wait when it is blocked until it is accepted is also evaluated. Analytical results are compared with simulation results.

3.2 Overview of the ABT Protocol

There are two variants of the protocol. The first, called ABT with Delayed Transmission (ABT/DT), is intended to multiplex the so called Stepwise Variable Bit Rate Sources. These sources are expected to have a stepwise need of bandwidth. However there is a restriction on the sources which must tolerate a delay in the negotiation of an increase of bandwidth. Many data communications are typical examples of such sources.

Basically the ABT/DT works as follows. When a source wants an increase of bandwidth (for example, when it wants to transfer a burst), it sends a Request to the so called ABT Control Unit, situated at the ingress node. This Request is forwarded to the first switching element of the link, which checks whether it can allocate the increase of bandwidth or not. If it has enough bandwidth, the Request is forwarded to the next switching element and so on until it reaches the egress node. Eventually the egress node will send an acknowledgment back to the ABT Unit and the source will be allowed to transfer the burst. The time passed from the ABT Unit sending the request until receiving the acknowledgment is called the Round Trip Time. Note that during this time the switching elements have allocated bandwidth for the source, but the transmission has not started yet. Therefore this time is an overhead introduced by the protocol.

If a switching element is not able to allocate the requested increase of bandwidth, it discards the Request, and by a time-out mechanism the allocated resources are reset to their previous state. In this case the ABT Unit makes successive reattempts until the increase of bandwidth is accepted. The source indicates the ABT Unit when an accepted burst is already transferred in order to release the allocated bandwidth.

The other variant, called ABT with immediate transmission (FRT/IT), is intended for sources more sensitive to a time delay. In this case the source transfers the burst immediately after the reservation request. If the reservation fails in any of the nodes, the whole burst is discarded.

3.3 Model Description and Analysis

In the analysis an isolated node is considered. Sources are assumed ON-OFF with exponential ON and OFF time distribution (burst-silence model). The parameters of the sources are the bit rate within a burst period Λ; the mean burst duration t^{on} and the mean silence duration t^{off}.

In the model N_l identical sources (in this chapter these are called *ltype* sources) with parameters Λ_l, t_l^{on} and t_l^{off}, are multiplexed with another different source (called *htype* source) with parameters Λ_h, t_h^{on} and t_h^{off}.

Being all time intervals exponentially distributed, the activation rate α of a source is given by $\alpha = 1/t^{off}$. Let the service time be the time that a node allocates bandwidth for a non blocked source. Clearly, for the ABT/IT the mean service time is the mean burst duration t^{on}. For the ABT/DT a non-blocked source has to wait a deterministic time equal to the round trip time t_{rt} before transferring a burst, so the mean service time is given by $t^{on} + t_{rt}$. However, in this analysis the influence of the round trip time time is not studied, so it is assumed to be zero. Therefore the service rate μ of a source is given by $\mu = 1/t^{on}$. Assuming $t_{rt} = 0$ the model makes no distinction between the ABT/DT and ABT/IT. Refer to [26] for a contrast of both variants of the protocol.

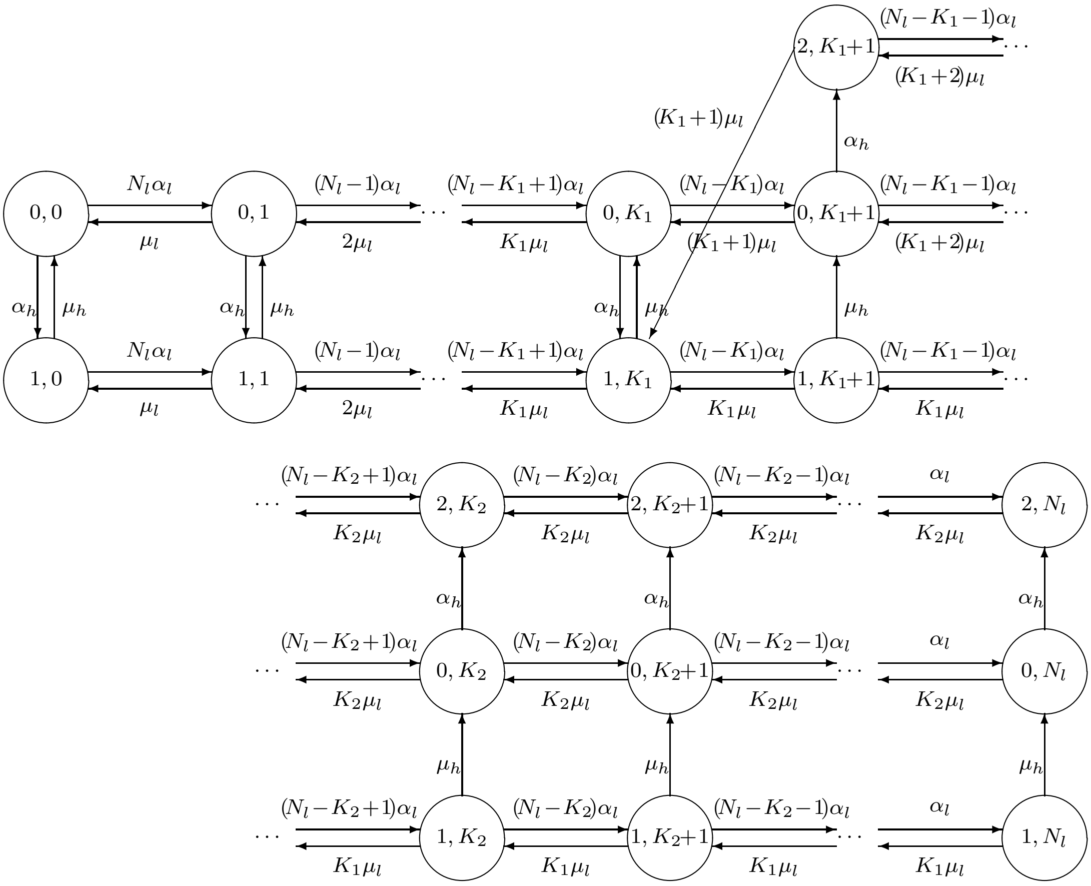


Figure 3.1: State-transition diagram assuming zero time between reattempts.

3.3.1 Approximation by Zero Time Between Reattempts

In this approximation we assume that when a burst is blocked, the time between the successive requests that are made until the burst is accepted is zero. This is equivalent to considering a blocked burst being kept in a queue until there is enough bandwidth left by the other sources in the link.

Let K_1 be the maximum number of ltype sources that can be simultaneously transferring a burst without exceeding the link capacity, when the htype source is also transferring a burst. Let K_2 be the same, but when the htype source is silent or blocked. Let us further suppose that the htype source transmits at a higher rate than the ltype source such that $K_2 > K1 + 1$. In this case when an ltype and htype sources are blocked, the ltype source will be accepted first (i.e. the htype source does not see a FIFO queue). Clearly, if the link capacity is C:

$$K_1 = \left\lfloor \frac{C - \Lambda_h}{\Lambda_l} \right\rfloor \tag{3.1}$$

$$K_2 = \left\lfloor \frac{C}{\Lambda_l} \right\rfloor \tag{3.2}$$

With these assumptions an isolated node can be described by the Markov chain of figure 3.1 with state space $\{(i, j) \ : \ i = 0, 1, 2 \ ; \ 0 \leq j \leq N_l\}$, where j is the number of ltype active sources (transferring or blocked) while the htype source is silent ($i = 0$), transferring a burst ($i = 1$) or blocked ($i = 2$). This Markov chain does not have a product form solution for the stationary probabilities π_{ij}, so they have to be calculated numerically solving the global balance equations.

The ltype and htype source blocking probability (P_l and P_h) can be obtained from the stationary probabilities π_{ij}. The blocking probability is given by the probability that an arriving burst is blocked, divided by the probability of a burst arrival. Thus:

$$P_h = \frac{\sum_{j=K_1+1}^{N_l} \pi_{0j}}{\sum_{j=0}^{N_l} \pi_{0j}} \tag{3.3}$$

$$P_l = \frac{\sum_{j=K_2}^{N_l-1} (N_l - j)\pi_{0j} + \sum_{j=K_1}^{N_l-1} (N_l - j)\pi_{1j} + \sum_{j=K_2}^{N_l-1} (N_l - j)\pi_{2j}}{\sum_{j=0}^{N_l-1} (N_l - j)\pi_{0j} + \sum_{j=0}^{N_l-1} (N_l - j)\pi_{1j} + \sum_{j=K_1+1}^{N_l-1} (N_l - j)\pi_{2j}} \tag{3.4}$$

3.3.2 Approximation by Identically Reattempt and OFF Time Distribution

In this approximation we assume that when a burst is blocked, the time between the successive requests that are made until the burst is accepted is exponentially distributed with a mean equal to the OFF time distribution, i.e. we assume an identically reattempt and OFF time distribution. This is equivalent to considering that a blocked burst is lost.

Let K_1 and K_2 be the same as in the previous section. Because a blocked burst can be considered as lost, with this approach an isolated node can be described by the Markov chain with state space $\{(i, j) \ : \ i = 0, 1 \ ; \ 0 \leq j \leq K_2\}$ of figure 3.2, where j is the number of ltype sources transferring a burst while the htype source is silent ($i = 0$) or transferring a burst ($i = 1$). The stationary probabilities π_{ij} of the Markov chain has a straightforward product form solution given by:

$$\pi_{ij} = \frac{1}{G} \begin{pmatrix} N_l \\ j \end{pmatrix} \rho_h^i \, \rho_l^j \tag{3.5}$$

where G is the normalization constant, $\rho_l = \mu_l/\alpha_l$ and $\rho_h = \mu_h/\alpha_h$. Note that considering more than one htype source or even considering more than two types of sources, a product form solution would still apply.

In this model no distinction is made between a burst or a reattempt arrival. So the blocking probability is calculated as the probability that a burst or a reattempt arrival is blocked, divided by the probability of a burst or a reattempt arrival. Such blocking probability for the ltype and htype sources (P_l and P_h) is given by:

$$P_h = \frac{\sum_{j=K_1+1}^{K_2} \pi_{0j}}{\sum_{j=0}^{K_2} \pi_{0j}} \tag{3.6}$$

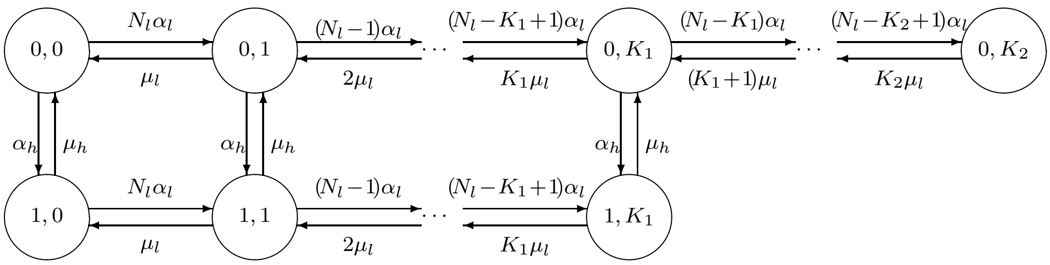


Figure 3.2: State-transition diagram assuming identically reattempt and OFF time distribution.

$$P_l = \frac{(N_l - K_2)\pi_{0K_2} + (N_l - K_1)\pi_{1K_1}}{\sum_{j=0}^{K_2}(N_l - j)\pi_{0j} + \sum_{j=0}^{K_1}(N_l - j)\pi_{1j}} \tag{3.7}$$

Note that in the previous section the reattempts to calculate the blocking probability are not counted (considering a zero time between reattempts implies considering ∞ reattempts after a blocked burst). If the reattempt time is not zero, the following relation applies for the P_l^{total} and P_l^{init} blocking probabilities of an ltype source, calculated counting and not counting the reattempts respectively. Let $\bar{r}_l$ be the mean number of reattempts that a blocked burst of an ltype source do until it is accepted. It can be derived that:

$$P_l^{init} = \frac{P_l^{total}}{\bar{r}_l\,(1 - P_l^{total})} \tag{3.8}$$

Obviously, an analogous relation holds for the htype source. If the blocking probability is small and the reattempt time is high enough such that $\bar{r} \approx 1$ (i.e. a blocked burst is almost always accepted at the first reattempt), $P^{init} \approx P^{total}$. These conditions are foreseeable in the approximation by identically reattempt and OFF time distribution. So, this approximation can be used to asses P_h^{init} and P_l^{init} from the probabilities calculated with equations (3.6) and (3.7).

3.3.3 Blocking Time in the Approximation by Zero Time Between Reattempts

In this section we calculate the time that an arriving burst that is blocked has to wait until it is eventually accepted (this is referred to as the *blocking time*). This time is calculated assuming the approximation by zero time between reattempts, so the referred states are those of figure 3.1. The approximation by identically reattempt and OFF time distribution is not used to assess the blocking time, because in general it would be inaccurate.

Let T_h and T_l be the blocking time of an htype source and ltype source respectively. Let $B_{ij} = (i, j)$ be the entering state resulting from the blocking transition. Clearly:

$$P(T_h \leq x) = \sum_{j=K_1+1}^{N_l} P(T_h \leq x|B_{2j})\,P(B_{2j})\,, \tag{3.9}$$

$$P(B_{2j}) = \frac{\pi_{0j}}{\sum_{k=K_1+1}^{N_l} \pi_{0k}} \tag{3.10}$$

and:

$$P(T_l \leq x) = \sum_{\forall B_{ij}} P(T_l \leq x | B_{ij})\, P(B_{ij})\,, \tag{3.11}$$

$$P(B_{ij}) = \frac{(N_l - j + 1)\pi_{ij-1}}{\displaystyle\sum_{k=K_2}^{N_l-1} (N_l - k)\pi_{0k} + \sum_{k=K_1}^{N_l-1} (N_l - k)\pi_{1k} + \sum_{k=K_2}^{N_l-1} (N_l - k)\pi_{2k}} \tag{3.12}$$

$P(T_h \leq x|B_{ij})$ is the distribution of the time that a blocked burst of an htype source has to wait until it is accepted, when the entering state in the blocking transition is B_{ij}. $P(T_l \leq x|B_{2j})$ is the same for an ltype source. Formulas for these probabilities are derived in appendixes 3.A and 3.B of this chapter.

3.4 Numerical Results

In this section we present a numerical study of the ABT fairness using the models described above. The fairness of the protocol is evaluated in terms of the burst blocking probability and the mean blocking time. Blocking time is specially important when using the ABT/IT scheme in which the sources are supposed to be time sensitive. Analytical and simulation results are also compared.

Figures 3.3 and 3.4 (model parameters are summarized in table 3.1) plot the blocking probability and the mean blocking time of the two source types considered, when the htype source varies the mean burst duration (i.e. the mean ON time t_h^{on}) [1]. When varying t_h^{on} from 0 (the source is always silent) to ∞ (the source is always active), the blocking probability of the ltype sources will increase from the one obtained when sharing a link of capacity varying from C to $C - \Lambda_h$. Figure 3.3 shows that the blocking probability of the ltype sources increases within these limits, while the blocking probability of the htype source remains constant. The blocking probability is assessed using the approximation by zero time between reattempts (section 3.3.1) [2], and the approximation by identical reattempt and OFF time distribution (section 3.3.2).

Figures 3.5 and 3.6 plot the blocking probability and the mean blocking time of the two source types, when the htype source varies the bitrate within a burst period. Each time that the htype source bitrate reaches a multiple of the ltype source bitrate, there is a decrement on the maximum number of sources that can be simultaneously transferring a burst. This causes an increasing step on the blocking probability and the blocking time.

Table 3.2 compares analytical and simulation results (given with 95% confidence intervals). To calculate the blocking probabilities in the simulation, the reattempts have not been counted in order to compare with the approximation by zero time between reattempts (these probabilities are referred to as “init.” in the table), and have been counted to compare with the approximation by identically reattempt and OFF time distribution (referred to as “tot.” in the table, cfr. section 3.3.2). Increasing the reattempt time decreases the blocking probability. So, the first approximation can be considered as an upper bound for the “init.” probabilities, and, for a

[1] Note that the burstiness, defined as $b = (t^{on} + t^{off})/t^{on}$ is a decreasing function with increasing t^{on}.

[2] To calculate the stationary probabilities using this approximation, the global balance equations have been solved using a Gaussian elimination method.

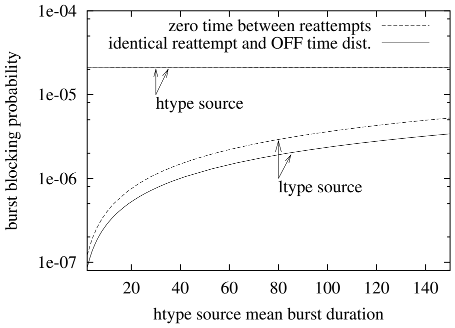


Figure 3.3: Influence of the htype source mean burst duration on the blocking probability.

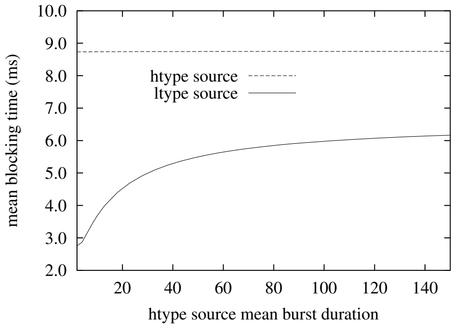


Figure 3.4: Influence of the htype source mean burst duration on the mean blocking time.

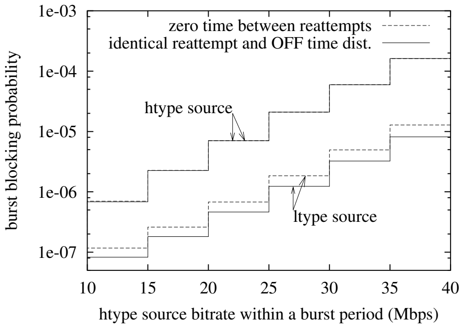


Figure 3.5: Influence of the htype source bit rate within a burst period on the blocking probability.

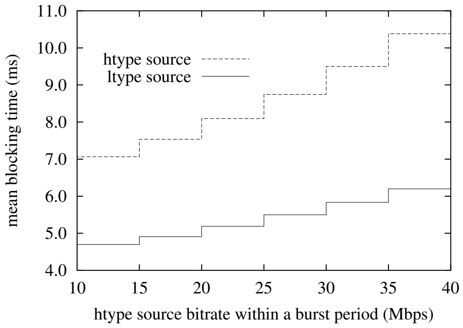


Figure 3.6: Influence of the htype source bit rate within a burst period on the mean blocking time.

reattempt time lower than the mean OFF time, the second approximation can be considered as a lower bound for the "tot." probabilities.

A deterministic and an exponentially distributed reattempt time has been considered in the simulation. It can be seen that the exponentially distributed approximation for the reattempt time gives accurate results for the blocking probabilities, but the blocking time. Simulation results show that the mean blocking time increases rapidly with increasing the reattempt time.

3.5 Conclusions

In this chapter we have analyzed the behavior of the ABT when different source types are multiplexed together. We have considered the case in which a set of identical sources is multiplexed with another one of higher bitrate. To assess the blocking probability two approximations have been considered. In the first one we assume that the reattempt time is zero and in the second one that it is identically distributed to the OFF time. To calculate the stationary state probabilities with the first approach the balanced global equations have to be solved, while in the second approach they have a simple product form solution. The mean blocking time has been also calculated assuming the first approach.

link capacity 150 Mbps	htype source				ltype source				
	bit rate (Mbps)	t_h^{on} (ms)	t_h^{off} (ms)	burstiness	bit rate (Mbps)	t_l^{on} (ms)	t_l^{off} (ms)	burstiness	num. of sources
fig. 3 & 4	30	0∼150	1400	$\infty \sim 10.3$	5	100	1400	15	150
fig. 5 & 6	10∼40	50	1400	29	5	100	1400	15	150
table 3.2	30	50	900	19	5	100	900	10	150

Table 3.1: Model parameters.

	Analytical		Simulation					
	Zero time between reatt.	Id. reatt. and OFF time dist.	Reattempt time					
			5 ms		20 ms		50 ms	
			Exp. dist.	Det.	Exp. dist.	Det.	Exp. dist.	Det.
P_h init.	$7.63\ 10^{-3}$		$7.00\ 10^{-3}$ $\pm 2.83\ 10^{-4}$	$6.86\ 10^{-3}$ $\pm 4.30\ 10^{-4}$	$7.31\ 10^{-3}$ $\pm 5.33\ 10^{-4}$	$7.28\ 10^{-3}$ $\pm 6.46\ 10^{-4}$	$7.00\ 10^{-3}$ $\pm 2.49\ 10^{-4}$	$7.16\ 10^{-3}$ $\pm 8.72\ 10^{-4}$
P_l init.	$8.22\ 10^{-4}$		$5.76\ 10^{-4}$ $\pm 1.19\ 10^{-5}$	$5.92\ 10^{-4}$ $\pm 1.78\ 10^{-5}$	$5.13\ 10^{-4}$ $\pm 4.89\ 10^{-5}$	$4.98\ 10^{-4}$ $\pm 5.28\ 10^{-5}$	$4.28\ 10^{-4}$ $\pm 1.10\ 10^{-5}$	$4.31\ 10^{-4}$ $\pm 4.37\ 10^{-5}$
P_h tot.		$7.57\ 10^{-3}$	$29.7\ 10^{-3}$ $\pm 1.65\ 10^{-3}$	$25.6\ 10^{-3}$ $\pm 1.75\ 10^{-3}$	$14.4\ 10^{-3}$ $\pm 1.33\ 10^{-3}$	$12.1\ 10^{-3}$ $\pm 1.34\ 10^{-3}$	$10.1\ 10^{-3}$ $\pm 0.35\ 10^{-3}$	$8.52\ 10^{-3}$ $\pm 1.12\ 10^{-3}$
P_l tot.		$4.15\ 10^{-4}$	$13.9\ 10^{-4}$ $\pm 2.32\ 10^{-5}$	$12.4\ 10^{-4}$ $\pm 4.29\ 10^{-5}$	$6.80\ 10^{-4}$ $\pm 6.36\ 10^{-5}$	$5.90\ 10^{-4}$ $\pm 6.57\ 10^{-5}$	$4.90\ 10^{-4}$ $\pm 1.09\ 10^{-5}$	$4.47\ 10^{-4}$ $\pm 4.63\ 10^{-5}$
T_h (ms)	15.24		22.0 ±0.32	19.13 ±0.48	40.8 ±0.95	33.6 ±0.88	73.2 ±1.23	59.9 ±0.92
T_l (ms)	6.822		12.3 ±0.07	10.51 ±0.16	28.0 ±0.24	23.7 ±0.22	57.5 ±0.41	51.8 ±0.16

Table 3.2: Comparison of analytical and simulation results.

The numerical study shows that there are not big differences between the blocking probabilities obtained with both approximations. The approximation of identical reattempt and OFF time distribution gives a much simpler way to compute the blocking probabilities and can be easily extended to more than one htype source or even more than two types of sources.

The results also show that when multiplexing different type of sources, blocking probability and blocking time depend on the source parameters. This can be interpreted as a lack of fairness, in the sense that they will have a different burst access. It is actually seen that an increase on the bitrate or the mean burst duration of a connection can result in a considerably increase of the blocking probability and blocking time of the other connections.

Appendixes

3.A Derivation of $P(T_h \leq x|B_{2j})$

In this appendix $P(T_h \leq x|B_{2j})$, $j \in \{K_1+1, \ldots , N_l\}$ of expression (3.9) is derived. If an htype and ltype sources are blocked, the ltype source will be accepted first (i.e. the htype source does not see a FIFO queue), so $P(T_h \leq x|B_{2j})$ is the distribution of the first passage time from the blocking state B_{2j} to the non blocking state $(1, K_1)$. To calculate this probability we follow the method described in [61].

To simplify the notation the states $(1, K_1), (2, K_1+1), \ldots , (2, N_l)$ will be referred to as E_j , $j = K_1, K_1+1, \ldots , N_l$. The following events are defined:

$T(j, j-r)$ = first passage time from state E_j to state E_{j-r}
$V(j, j-r)$ = number of transitions involved in $T(j, j-r)$

and their joint probability $G_j^{(r)}(x, k) = P\{T(j, j-r) \leq x, V(j, j-r) = k\}$. The probability we are looking for is given by:

$$P(T_h \leq x|B_{2j}) = \sum_{k=1}^{\infty} G_j^{(j-K_1)}(x, k) \tag{3.13}$$

Now $G_j^{(r)}(x, k)$ is computed. To simplify the notation, in case of one state transition we write $G_j(x, k) = G_j^{(1)}(x, k)$. Let q_j^+ be the transition rate from the state E_j to the state E_{j+1}; q_j^- the transition rate from the state E_j to the state E_{j-1}; and q_j the self state transition rate (cfr. figure 3.1), i.e.:

$$\begin{aligned} q_j^+ &= (N_l - j)\,\alpha_l \\ q_j^- &= \begin{cases} j\,\mu_l\ , & j \leq K_2 \\ K_2\,\mu_l\ , & j > K_2 \end{cases} \\ q_j &= q_j^+ + q_j^- \end{aligned}$$

We define the one state forward and backward transition probabilities: $A_j^-(x) = P\{T(j, j-1) \leq x, V(j, j-1) = 1\}$ and $A_j^+(x) = P\{T(j, j+1) \leq x, V(j, j+1) = 1\}$. We have:

$$\begin{aligned} A_j^-(x) &= (1 - e^{-q_j x})\frac{q_j^-}{q_j}\ ,\ j = K_1+1, \ldots , N_l \\ A_j^+(x) &= (1 - e^{-q_j x})\frac{q_j^+}{q_j}\ ,\ j = K_1, \ldots , N_l - 1 \end{aligned} \tag{3.14}$$

yielding:

$$G_j(x, k) = \begin{cases} A_j^-(x)\,, & k = 1 \\ 0\,, & k = 2n \\ \displaystyle\sum_{l=1}^{n} A_j^+(\cdot) * G_{j+1}(\cdot, 2(n-l)+1) * G_j(x, 2l-1)\,, & k = 2n+1 \end{cases} \tag{3.15}$$

and:

$$G_j^{(r)}(x,k) = \sum_{k_1+\cdots+k_r=k} G_j(\cdot,k_1) * G_{j-1}(\cdot,k_1) * \cdots * G_{j-r+1}(x,k_r) \tag{3.16}$$

where $*$ is the convolution of the distribution functions (i.e. $F_1(\cdot)*F_2(x) = \int_{-\infty}^{\infty} F_1(x-\lambda)\, dF_2(\lambda)$).

From (3.15) we derive the following recursive equation for the joint transform $\tilde{G}_j(s,z) = \sum_{k=0}^{\infty} \int_0^{\infty} e^{-sx} z^k \, dG_j(x,k)$:

$$\tilde{G}_j(s,z) = \frac{z\, A_j^-(s)}{1 - z\, A_j^+(s)\, \tilde{G}_{j+1}(s,z)} \tag{3.17}$$

and:

$$\tilde{G}_j^{(r)}(s,z) = \sum_{k=0}^{\infty} z^k \sum_{k_1+\cdots+k_r=k} G_j(s,k_1)\, G_{j-1}(s,k_1) \cdots G_{j-r+1}(s,k_r) = \\ \tilde{G}_j(s,z)\, \tilde{G}_{j-1}(s,z) \cdots \tilde{G}_{j-r+1}(s,z) \tag{3.18}$$

where $A_j^-(s)$, $A_j^+(s)$ and $G_j(s,k)$ are the Laplace-Stieltjes transforms of the distribution functions (i.e. $F(s) = \int_0^{\infty} e^{-sx} \, dF(x)$). From (3.14) we obtain:

$$\begin{aligned} A_j^-(s) &= \frac{q_j^-}{s+q_j}\,,\ j = K_1+1,\dots,N_l \\ A_j^+(s) &= \frac{q_j^+}{s+q_j}\,,\ j = K_1,\dots,N_l-1 \end{aligned} \tag{3.19}$$

Substitution into (3.17) yields:

$$\tilde{G}_j(s,z) = \frac{z\, q_j^-}{s + q_j^+ - z\, q_j^- \tilde{G}_{j+1}(s,z)}\,,\ j = K_1+1,\dots,N_l-1 \tag{3.20}$$

$$\tilde{G}_{N_l}(s,z) = z\, \frac{q_{N_l}^-}{s + q_{N_l}^-} \tag{3.21}$$

Substituting recursively (3.21) into (3.20), and then into (3.18) we obtain $\tilde{G}_j^{(r)}(s,z)$, $j = K_1 + 1,\dots,N_l$. Finally, from (3.13) we derive that $G_j^{(j-K_1)}(s,z)_{z=1}$ is the Laplace-Stieltjes transform of $P(T_h \le x | B_{2j})$. Inverting it and substituting into (3.9) we obtain the distribution of the blocking time T_h. This is rather arduous, but from the previous equations a straightforward formula for the mean blocking time $\bar{T}_h$ can be derived.

Define $\bar{T}_j^{(j-K_1)} = \mathrm{E}[x|B_j]$, i.e. the mean first passage time from the state E_j to the state E_{k1}. Define also $\bar{T}_j$ as the mean first passage time from the state E_j to the state E_{j-1}. Clearly $\bar{T}_j^{(j-K_1)} = -\left.\frac{\partial}{\partial s}\tilde{G}_j^{(j-K_1)}(s,z)\right|_{s=0,\,z=1}$, $\bar{T}_j = -\left.\frac{\partial}{\partial s}\tilde{G}_j(s,z)\right|_{s=0,\,z=1}$.

From (3.20), (3.21) and (3.18), and since $\tilde{G}_j(s,z)_{s=0,\,z=1} = 1$ we obtain:

$$\bar{T}_j = \frac{1 + q_j^+ \, \bar{T}_{j+1}}{q_j^-} \tag{3.22}$$

$$\bar{T}_{N_l} = \frac{1}{q_{N_l}^-} \tag{3.23}$$

$$\bar{T}_j^{(j-K_1)} = \sum_{k=K_1+1}^{j} \bar{T}_k \tag{3.24}$$

Substituting recursively (3.23) into (3.22), and then into (3.24), $\bar{T}_j^{(j-K_1)}$ is computed. Finally from (3.9) the mean blocking time is obtained as:

$$\bar{T}_h = \sum_{j=K_1+1}^{N_l} \bar{T}_j^{(j-K_1)} P(B_{2j}) \tag{3.25}$$

3.B Derivation of $P(T_l \leq x|B_{ij})$

In this appendix we derive the $P(T_l \leq x|B_{ij})$ of expression (3.11). To calculate this probability the following cases are considered:

1. $B_{ij} \in \{(2, K_2+1), \ldots, (2, N_l), (0, K_2+1), \ldots, (0, N_l)\}$

In this case the ltype source is blocked while the htype source is silent or blocked. Being $B_{ij} = (i,j)$ the state resulting from the blocking transition, the source will find $j - K_2 - 1$ ltype sources already blocked and it will have to wait until $j - K_2$ ltype sources are served (a source is said to be served when one of the K_2 bursts being transferred ends). Let $S_l^{(n)}$ be the service time of n ltype sources and $F_{S_l}^{(n)}(x)$ its distribution. Clearly $F_{S_l}^{(1)}(x) = 1 - e^{-K_2\mu_l\, x}$ and $F_{S_l}^{(n)}(x) = F_{S_l}^{(1)}(\cdot) * \overset{n}{\ldots} * F_{S_l}^{(1)}(x)$. Let $F_{T_l|B_{ij}}(s)$ be the Laplace-Stieltjes transform of $P(T_l \leq x|B_{ij})$. The following relation holds:

$$F_{T_l|B_{ij}}(s) = F_{S_l}^{(j-K_2)}(s) = \frac{(K_2\mu_l)^{j-K_2}}{(s + K_2\mu_l)^{j-K_2}} \tag{3.26}$$

2. $B_{ij} \in \{(1, K_1+1), \ldots, (1, K_2)\}$

In this case the ltype source is blocked while the htype source is transferring a burst, but the maximum of active ltype sources is K_2. Being $B_{ij} = (i,j)$ the state resulting from the blocking transition, the source will find $j - K_1 - 1$ ltype sources already blocked. So to be accepted it will have to wait until the htype source is served or until $j - K_1$ ltype sources are served. Let S_h be the service time of the htype source and $S_l^{(n)}$ the service time of n ltype sources. Let $F_{S_h}(x)$ and $F_{S_l}^{(n)}(x)$ be their distribution. Clearly $F_{S_h}(x) = 1 - e^{-\mu_h\, x}$ and $F_{S_l}^{(n)}(x)$ is the same as in the previous case, but changing K_2 for K_1. Therefore:

$$\begin{aligned} P(T_l \leq x|B_{ij}) = 1 - P(S_h > x)\, P(S_l^{(j-K_1)} > x) = \\ 1-(1-F_{S_h}(x))(1-F_{S_l}^{(j-K_1)}(x)) = 1-(1-F_{S_l}^{(j-K_1)}(x))\, e^{-\mu_h\, x} \end{aligned} \tag{3.27}$$

the Laplace-Stieltjes transform of the previous equation is:

$$\begin{aligned} F_{T_l|B_{ij}}(s) = s\int_0^\infty \left(e^{-sx} - (1-F_{S_l}^{(j-K_1)}(x))e^{-(s+\mu_l)x}\right) dx = \\ 1 - \frac{s}{s+\mu_h}\left[1 - \left(\frac{K_1\mu_l}{s+\mu_h+K_1\mu_l}\right)^{j-K_1}\right] \end{aligned} \tag{3.28}$$

3. $B_{ij} \in \{(1, K_2+1), \dots, (1, N_l)\}$

In this case the ltype source is blocked while the htype source is transferring a burst, but there are more than K_2 active ltype sources. So although the htype source is served, the ltype source can still remain blocked. Let S_h be the service time of the htype source and $S_l^{(n)}$ the service time of n ltype sources while the htype is being served. Let us consider the density of the blocking time. For notation convenience we define $P_{ij}(T_l = x) = P(T_l = x|B_{ij})$. Clearly:

$$\begin{aligned} P_{ij}(T_l = x) = P_{ij}(T_l = x,\, S_h < S_l^{(1)}) + \\ \sum_{k=1}^{j-K_2-1} P_{ij}(T_l = x,\, S_l^{(k)} < S_h < S_l^{(k+1)}) + P_{ij}(T_l = x,\, S_h > S_l^{(j-K_2)}) \end{aligned} \tag{3.29}$$

After some computation, the Laplace transform of the previous expression yields:

$$\begin{aligned} F_{T_l|B_{ij}}(s) = \sum_{k=0}^{j-K_2-1} \mu_h \left(\frac{K_2\mu_l}{s+K_2\mu_l}\right)^{j-k-K_2} \frac{(K_1\mu_l)^k}{(s+\mu_h+K_1\mu_l)^{k+1}} + \\ \left[1 - \frac{s}{s+\mu_h}\left[1-\left(\frac{K_1\mu_l}{s+\mu_h+K_1\mu_l}\right)^{K_2-K_1}\right]\right]\left(\frac{K_1\mu_l}{s+\mu_h+K_1\mu_l}\right)^{j-K_2} \end{aligned} \tag{3.30}$$

Inversion of (3.26), (3.28) and (3.30), and substitution into (3.11) yields the distribution of the blocking time T_l. Differentiating these equations the mean blocking time is calculated as:

$$\bar{T_l} = \sum_{\forall B_{ij}} -\frac{d}{ds}\, F_{T_l|B_{ij}}(s)\Big|_{s=0} P(B_{ij}) \tag{3.31}$$

Chapter 4

The Available Bit Rate (ABR) Service

4.1 Introduction

This chapter describes the Available Bit Rate Service Category (ABR) principles and rules as defined by the ATM Forum [4].

As mentioned in section 2.5.4, in the ABR Service the sources adapt their transmission rate to the available bandwidth left by the guaranteed rate traffic. Rate however, is not arbitrary adjusted, but a commitment is made by the network that bandwidth received by sources sharing the same link is fairly apportioned.

The transmission rate of each connection is controlled by means of special cells called Resource Management cells (RM-cells). RM-cells are transmitted embedded in the Data-Cell flow from the Source End System (SES) to the Destination End System (DES) (see figure 4.1). The DES "turns around" the RM-Cells, which return to the SES along the same path. In order to convey cell rate adjustments to the sources, the switches modify the contents of some fields of the RM-cells.

The standard specifies the source and destination behavior and several fields of the RM-cells that can be used by the switches to control source rates. The algorithm applied by the Switches to make use of these fields is implementation specific. In the following the RM-cell format, the rules specifying the source behavior and the feedback mechanisms available for the switches are described.

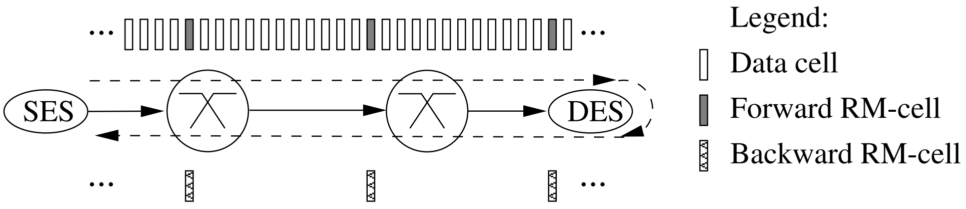


Figure 4.1: RM-cell flow in the ABR Service Category

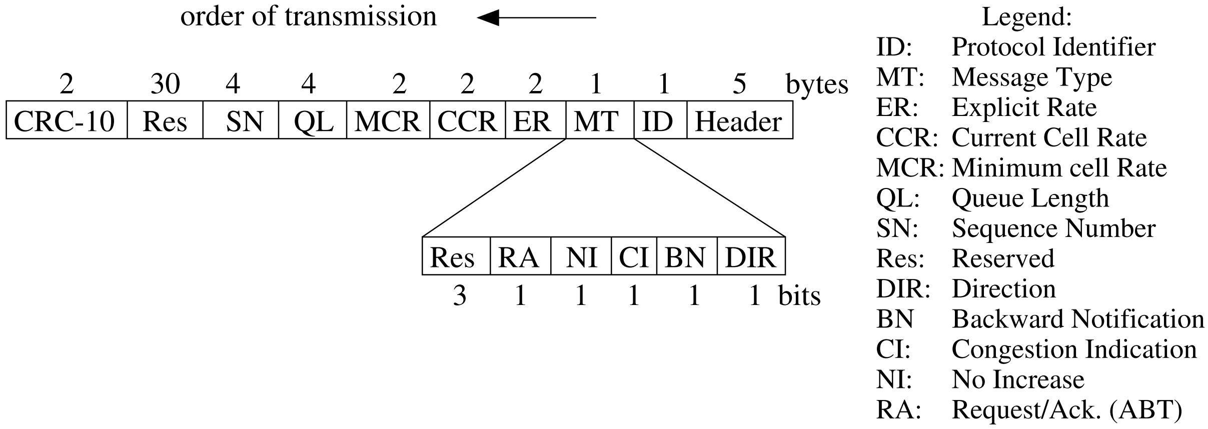


Figure 4.2: RM-cell format.

4.2 RM Cell Format

Figure 4.2 shows the fields of the RM-cells. The RM-cell header is the standard ATM-cell header shown in figure 2.3 with the payload type indicator bits equal to 110 (cfr. table 2.2). Additionally, a virtual path connection must set the VCI to 6. The other fields of the RM-cells are described in the following.

ID is the protocol identifier. ITU-T has assigned ID=1 for ABR. Other protocols using RM-cells, as ABT, use different values.

MT is called the message type field and contains the following bits:

- **DIR** indicates if it is a forward RM-cell (DIR=0) or a backward RM-cell (DIR=1).
- **BN** indicates if a backward RM-cell has been generated by a SES (BN=0), or by a switch or DES (BN=1).
- **CI** is called the *Congestion Indication* bit. When it is set the source decreases the transmission rate.
- **NI** is called the *No Increase* bit. Switches may set this bit to avoid the sources increase the transmission rate.
- **RA** bit is not used by ABR. This is defined for compatibility with ABT.

ER is called the *Explicit Rate* field. This is used by the switches to explicitly set the source transmission rate.

CCR is called the *Current Cell Rate* field. This field is set by the SES to the ACR (see section 4.3) and may be used by the switches as an estimation of the source transmission rate.

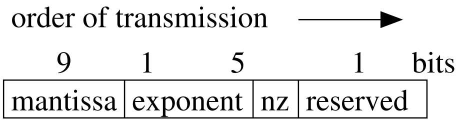


Figure 4.3: Rate format used in the RM-cell fields.

MCR is set by the SES to the Minimum Cell Rate negotiated at the connection set up.

The ER, CCR and MCR fields specify the transmission rate using the format shown in figure 4.3. The rate, in cells per second, specified by this unsigned floating point representation is given by:

$$\text{Rate} = \text{nz} \cdot \left(1 + \frac{\text{mantissa}}{512}\right) \cdot 2^{\text{exponent}} \tag{4.1}$$

This coding allows the representation of a maximum value of $(1+511/512)\cdot 2^{31} = 4,290,772,992$ cells/s. However, during the connection setup the rates are negotiated using a 24 bits integer format, which limits the maximum value to $2^{24} = 16,777,216$ cells/s.

QL and SN are not used by the ATM Forum ABR. These fields have been defined for compatibility with the ITU-T specification (see [35]).

4.3 ABR Source Behavior

The ATM Forum [4] describes the behavior of an ABR source by means of a set of Source End System (SES) and the Destination End System (DES) rules. An ABR connection is always bi-directional and the sources located at both termination points must implement the SES and DES rules. The parameters involved in these rules are summarized in table 4.1.

The ACR is a parameter kept by the source which fixes the maximum rate at which cells can be scheduled for transmission. This parameter is adjusted according to the feedback conveyed by the backward RM-cells. The PCR and MCR are an upper and lower bound of the ACR. Therefore, by means of the PCR the source negotiates the maximum rate at which it may be allowed to send cells, and the MCR is a minimum guaranteed rate.

Some of these parameters (PCR, MCR, ICR, RIF, RDF, TBE, FRTT) are signaled and negotiated during the connection setup. The parameters indicated in table 4.2 are optionally signaled. In case they are not specified, the default value shown in table 4.2 is used. The parameters Mrm and TCR are not signaled since they have a fixed value. Finally, two parameters (CRM and ICR) are computed or updated upon completion of the call setup as:

$$\text{CRM} = \left\lceil \frac{\text{TBE}}{\text{Nrm}} \right\rceil \tag{4.2}$$

$$\text{ICR} = \min\left(\text{ICR}, \frac{\text{TBE}}{\text{FRTT}}\right) \tag{4.3}$$

Acronym	Description	Range	Units
ACR	Allowed Cell Rate	$0 \sim 2^{24}$	cells/s
PCR	Peak Cell Rate	$0 - 2^{24}$	cells/s
MCR	Minimum Cell Rate	$0 \sim 2^{24}$	cells/s
ICR	Initial Cell Rate	$0 \sim 2^{24}$	cells/s
RIF	Rate Increase Factor	$1/2^{16} \sim 1/2^{0}$	
Nrm	Maximum distance, in cells, between forward RM-cells	$2^{1} \sim 2^{8}$	
Mrm	Controls RM-cell and data cell transmissions	2	
RDF	Rate Decrease Factor	$1/2^{16} \sim 1/2^{0}$	
CRM	Missing RM-cell count		cells
ADTF	ACR Decrease Factor	$0.01 \sim 10.23$	sec
Trm	Upper bound between forward RM-cell transmission	$100 \cdot 2^{-7} \sim 100 \cdot 2^{0}$	ms
FRTT	Fixed Round Trip Time	$0 \sim 16.7$	sec
TBE	Transient Buffer Exposure	$0 \sim 2^{24} - 1$	cells
CDF	Cutoff Decrease Factor	$1/2^{6} \sim 1/2^{0}$	
TCR	Tagged Cell Rate	10	cells/s

Table 4.1: Parameters involved in the ABR source behavior.

The following sections describe the SES and DES set of rules and the meaning of these parameters.

Parameter	Default value
Nrm	32
Trm	100
ADTF	0.5
CDF	1/16

Table 4.2: Optionally signaled ABR parameters.

4.3.1 Source End System Rules

SES is responsible of the following functions:

- Adjust the cell transmission rate according to the feedback conveyed by the backward RM-cells,
- compute the cell transmission time,
- generate forward RM-cells,

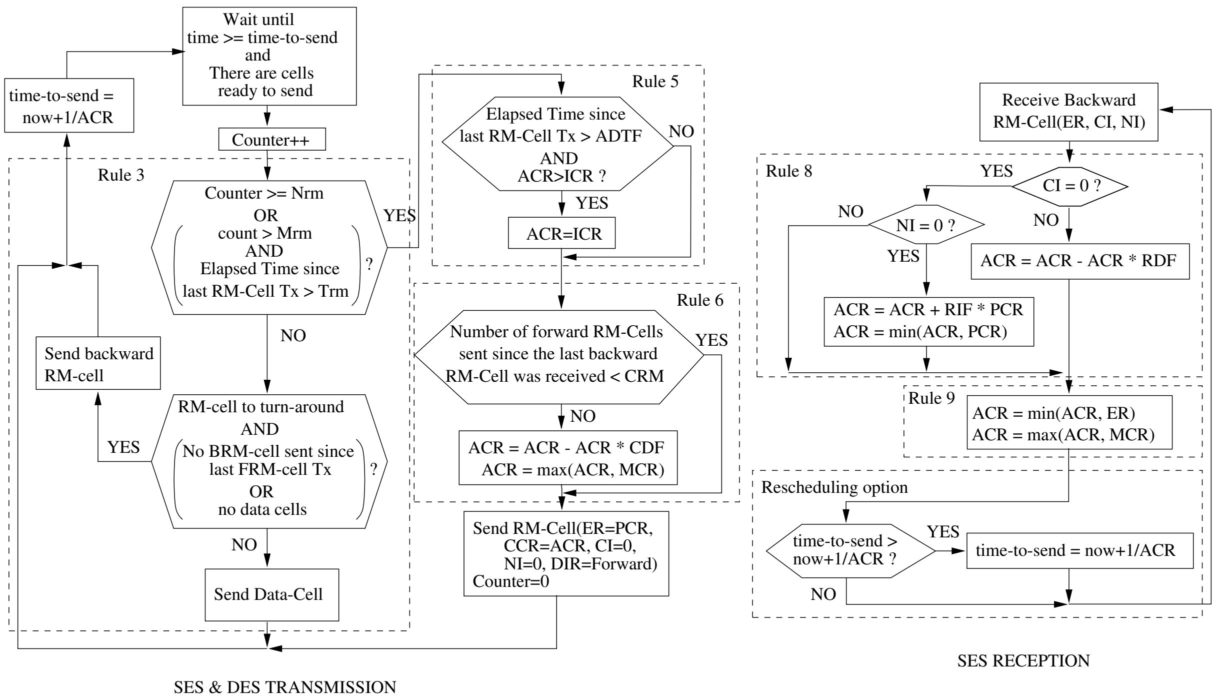


Figure 4.4: SES and DES behavior.

- "turn around" the forward RM-cells received from the remote SES, and send them as backward RM-cell,
- schedule the transmission of data cells, forward RM-cells and backward RM-cells.

As mentioned before, the ACR is the maximum rate at which cells may be scheduled for transmission. This rate includes the transmission of data cells, forward RM-cells and backward RM-cells. According to this definition, a source having the ACR decreased down to zero could be stopped. To avoid this problem the ATM Forum defines the *in-rate* and *out-of-rate* RM-cells. In-rate RM-cells are sent with the CLP bit (see section 2.3) set to 0, and out-of-rate with CLP=1. The out-of-rate RM-cells are not counted in the ACR. However, to avoid overloading the network the rate at which a source may generate out-of-rate RM-cells is limited by the TCR parameter (which has been fixed to 10 cells/s).

The ATM-Forum has specified the SES functions by means of 13 rules. These rules are discussed in the following. To better understand the rules, these have been summarized with the flow chart of figure 4.4.

SES rule 1: The value of the ACR must always satisfy: MCR $\leq$ ACR $\leq$ PCR

SES rule 2: After the connection setup the ACR must be set at most to the ICR and the first transmitted cell must be a forward RM-cell.

SES rule 3: This rule refers to the transmission scheduling of data cells, forward RM-cells and backward RM-cells. The rule states that after the first forward RM-cell transmission, the next in-rate cell shall be:

1. a forward RM-cell if since the last in-rate forward RM-cell was sent, either: i) at least Mrm in-rate cells have been sent and at least Trm time has elapsed, or ii) Nrm-1 in-rate cells have been sent.
2. a backward RM-cell if the former condition is not met, if any backward RM-cell is waiting for transmission and if either: i) no in-rate backward RM-cell has been sent since the last in-rate forward RM-cell or ii) no data cell is waiting for transmission.
3. a data cell if former conditions are not met and any data cell is waiting for transmission.

The flow chart of figure 4.4 depicts a possible implementation of the former conditions. Note that, normally, each Nrm-1 data cells a forward RM-cell will be sent. Just in case the ACR is low, a forward RM-cell will be sent if Trm time has elapsed since the previous RM-cell was sent. Trm as the default value of 100 ms (see table 4.2). The parameter Mrm (which has been fixed to 2) avoids a low rate source from sending only forward RM-cells at each transmission time opportunity.

SES rule 4: All cells transmitted according to previous rules shall be in-rate cells (and thus, will be transmitted with CLP=0).

SES rule 5: When a forward RM-cell is going to be transmitted, if the elapsed time since the last forward RM-cell transmission is higher than the ADTF parameter, the ACR must be reduced down (if higher) to the ICR.

This adjustment is intended to solve the problem referred to as *ACR retention* . This problem arises when a source starts a transmission after a long idle period. In this case the bandwidth may be given to other sources. Therefore, the source could start transmitting at the full ACR resulting in a harm for the network if the last computed ACR was too high.

SES rule 6: After the former rule, if a number $\geq$ CRM of forward RM-cells have been sent, the ACR must be reduced by at least $\mathrm{ACR} \cdot \mathrm{CDF}$ (see figure 4.4).

This rule is intended to protect the network in case of the backward RM-cell flow being blocked because of heavy congestion or a broken VC. Note that the SES could not receive feedback in these cases.

Since the rule can be triggered unnecessarily in case of a long round trip time between the SES and the DES, CRM is computed by means of equation (4.2). In this equation TBE determines the maximum number of cells that may transmitted by the source during the first round trip time, and thus, before the feedback control takes effect. Therefore, equation (4.2) estimates the number of forward RM-cells sent during this first round trip time.

SES rule 7: The SES must place the current ACR into the CCR field of the outgoing forward RM-cells. This field may be used by the switches as an estimation of the source rate in order to fairly divide the available bandwidth.

SES rule 8: This rule together with rule 9 specifies how the SES has to adjust the ACR according to the feedback conveyed by the backward RM-cells (see figure 4.4). Rule 8 states that upon reception of a backward RM-cell, the CI and NI bits shall be tested and the ACR adjusted according to the table 4.3.

CI	NI	ACR adjustment
0	0	ACR = min(ACR + RIF · PCR, PCR)
1	0/1	ACR = max(ACR - RDF · ACR, MCR)
0	1	ACR

Table 4.3: ACR adjustment according to CI and NI bits.

SES rule 9: After rule 8 is applied, rule 9 states that the ACR is to be adjusted to:

$$\mathrm{ACR} = \max(\min(\mathrm{ACR}, \mathrm{ER}), \mathrm{MCR})$$

That is, the ACR must be set at most to the contents of the ER field of the backward RM-cell, but not below the MCR.

After this final update of the ACR with the backward RM-cell contents, the ATM Forum allows the SES to perform a "Rescheduling option". This option consists of rescheduling a transmission time of a cell in order to take advantage of an increase in the ACR (see Appendix I of [4] for details).

SES rule 10: This rule states the values of the fields of a forward RM-cell generated by a SES (see section 4.2). These are shown in the following table:

field	ID	DIR	BN	CI	NI	ER	CCR	MCR
value	1	0	0	0	0	PCR	ACR	MCR

Table 4.4: Initial values of a forward RM-cell.

The bit RA and the fields QL and SN are either set to zero or in accordance to the ITU-T specification [35].

SES rule 11: As mentioned before, a SES may send out-of-rate forward RM-cells in some cases, e.g. if the ACR is zero these cells may be sent periodically looking for the opportunity to increase the current rate. This rule states that forward RM-cells cannot be sent at a rate greater than the TCR.

SES rule 12: This rule states that the SES has to reset the EFCI bit of the header from all outgoing data cells (see table 2.2 for a description of the EFCI bit).

SES rule 13: Remember from rule 5 that *ACR retention* may cause a harm for the network in case a source with a long ACR starts transmitting at the full ACR after a long idle period. Rule 5 intends to solve this problem by setting the ACR to the ICR if the time between two consecutive forward RM-cells is higher than the ADTF parameter.

Notice, however, that if a source becomes internally rate-limited but not idle, the ACR retention could also happen without being prevented by the ADTF mechanism. To avoid this the ATM Forum leaves as optional the implementation of *use-it-or-lose-it* policies. This use-it-or-lose-it behavior intends to maintain the ACR to a value that approximates the actual cell transmission rate of the SES.

4.3.2 Destination End System Rules

The main function of the DES is to *turn around* the forward RM-cells received from the SES, i.e. the transmission of a backward RM-cell in response to having received a forward RM-cell.

DES rule 1: The DES has to keep track of the EFCI bit conveyed by the data cells and store the last value as the *EFCI state* of the connection.

DES rule 2: On receiving a forward RM-cell, the DES should turn around the cell changing the fields of the RM-cell in the following way:

- The DIR bit is changed from *forward* (0) to *backward* (1).
- The BN bit is set to 0.
- If the *EFCI state* of the connection is set (see above), the CI bit has to be set and the EFCI state reset.
- A DES having internal congestion may reduce the ER and/or set the CI or the NI bits.

DES rule 3: If the ACR is low, the DES may receive forward RM-cells to turn around while a backward RM-cell is waiting for transmission. In order to send the most updated feedback, this rule states that in this case:

- It is recommended that the contents of the old backward RM-cell are overwritten by the contents of the new RM-cell to turn around.
- It is recommended that the old cell is either sent out-of-rate, discarded, or remain scheduled for in-rate transmission.
- The new cell has to be scheduled for in-rate transmission.

DES rule 4: Regardless of the alternative chosen when applying the previous rule, the contents of the older cell do not have to be transmitted after the contents of the new cell.

DES rule 5: This rule states that a DES may generate backward RM-cells without having to wait for a forward RM-cell to turn around. This may be useful in case the DES is congested and wants to reduce the SES transmission rate. These cells have to accomplish the following items:

- The rate of these backward RM-cells (including in-rate and out-of-rate) is limited to 10 cells/s.
- At least the CI or the NI bits have to be set (since these cells cannot increase the ACR of the SES).
- The DIR bit has to be set to backward (1) and the BN has to be set.

DES rule 6: When an out-of-rate forward RM-cell (with CLP=1) is turned around, it may be sent in-rate or out-of-rate.

4.4 Switch behavior

For ABR switches the ATM Forum has only specified the way feedback information is conveyed to the ABR sources. The exact algorithm used by the switches to perform the feedback updates is implementation specific. This makes possible a diversity of switches with different degrees of complexity and performance, maintaining compatibility with the standard. The ATM Forum has specified the way feedback information can be conveyed to the ABR sources by means of the following 5 rules:

Switch rule 1: This rule states that a switch shall implement at least one of the following methods to control congestion:

- *EFCI Marking* consists of setting the EFCI bit of the data cells headers;
- *Relative Rate Marking* consists of setting the CI or NI bit of forward and/or backward RM-Cells;
- *ER Marking* consists of reducing the ER of forward and/or backward RM-Cells and
- *VS/VD Control* consists of segmenting the control loop by implementing a *Virtual Destination* (VD), which behaves like a DES, and a *Virtual Source* (VS), which behaves like a SES (see figure 4.5). This method may be useful in order to avoid the low responsiveness that could arise in an ABR control loop with a long round trip time. Note however that the switch would require a queue for each connection having the VS/VD control.

Switch rule 2: In case of severe congestion a switch may want to reduce the source rate immediately. Similarly to DES rule 5, this rule allows the switch to generate backward RM-cells. The generation rate of these cells is limited to 10 cells/s for each connection. Furthermore, at least the CI or the NI bits have to be set, the DIR bit has to be set to backward (1) and the BN has to be set.

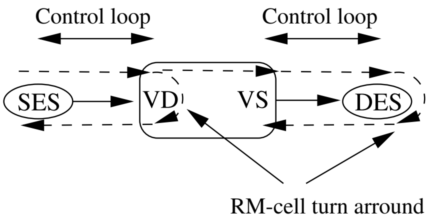


Figure 4.5: ABR switch with a VS/VD control.

Switch rule 3: This rule states that the RM-cell sequence with respect to the data cells may not be maintained. This may increase the feedback responsiveness in case of congestion. However, sequence integrity within the RM-cell stream must be maintained.

Switch rule 4: This rule specifies the fields that may be changed by an ABR switch. These are: the CI, NI, and ER according to rule 1. The RA bit and the QL and SN fields have to be set in accordance to ITU-T specification [35]. The MCR may be corrected to the negotiated MCR at the connection setup.

Switch rule 5: A switch may implement a use-it-or-lose-it policy in order the sources maintains the ACR value close to the actual transmission rate (see section 4.3.1, SES rule 13).

4.4.1 Fairness algorithm

The ATM Forum establishes that ABR switches should allocate the bandwidth among the sources following any fairness scheme [79]. Several fairness criteria have been proposed, although the algorithm is implementation specific. Depending on the switch mechanism, fairness is achieved with different degrees of accuracy and convergence speed. The Max-Min fairness criteria has been mainly used as a goal for switch mechanisms. This algorithm consists of computing the fair allocation B of the link. A contending source for the link which cannot transmit at B because is internally rate limited, or because is limited to a lower rate by another switch is said to be constrained. Unconstrained sources get the fair allocation B. B is computed as:

$$B = \frac{A - U}{N - N'} \tag{4.4}$$

where A is the available bandwidth on the link, N the total number of active sources sharing the link, U the sum of bandwidth of the constrained sources and N' its number.

4.4.2 Fairness index

In order to measure the degree of fairness achieved by the network R. Jain has proposed the following relation [39]:

$$\text{Fairness index} = \frac{\left(\sum_i x_i\right)^2}{n \sum_i x_i^2} \tag{4.5}$$

Where $x_i = \tilde{x}_i/\hat{x}_i$, being $\tilde{x}_i$ the transmission rate allocated for the source i and $\hat{x}_i$ the fair transmission rate to be allocated for that source.

Chapter 5

Switching Mechanisms for ABR

5.1 Introduction

As explained in section 4.4, for ABR switches the ATM Forum has only specified the way feedback information is conveyed to the ABR sources. The exact algorithm used by the switches to perform the feedback updates is implementation specific.

Switches using the *EFCI marking* or *relative rate marking* mechanisms (see section 4.4) are referred to as *binary switches*. It is foreseen that a first generation of ABR switches would adjust the transmission rate of the source by this simple binary indication [70]. Some examples of such switches are given in [80, 57].

A second generation of switches would perform a fairness algorithm to compute the transmission rate to allow to each connection and use the Explicit Rate (ER) field of the RM-cells to convey this rate to the sources. These are called *ER switches*. ER switch algorithms have been extensively studied in many contributions. Some of the proposed algorithms are based on the control theory [42, 32, 46, 11]. Other switch algorithms are directly derived from the fair bandwidth allocation criteria [68, 43, 77, 41]. It is worth mentioning that the ERICA switch algorithm [41] has become popular due to its simplicity and robustness, and has been used as a point of reference in many other proposals.

In this chapter three switch algorithms have been chosen to show the different degrees of performance and complexity that can be achieved. Further discussion and confront of ABR switches can be found in [2].

5.2 Description of Some Switch Algorithms

5.2.1 EFCI Switch

This switch mechanism was described in the earlier proposal of the Rate-Control algorithm [80]. It is the simplest switch mechanism. It monitors the queue length of the buffer and if it is larger than a threshold the switch is considered congested. During the congestion state the switch sets the EFCI bit of data cells to congested (EFCI=1). In order to reduce the feedback delay, a further proposal establishes setting CI=1 of backward RM-cell bit during the congestion state

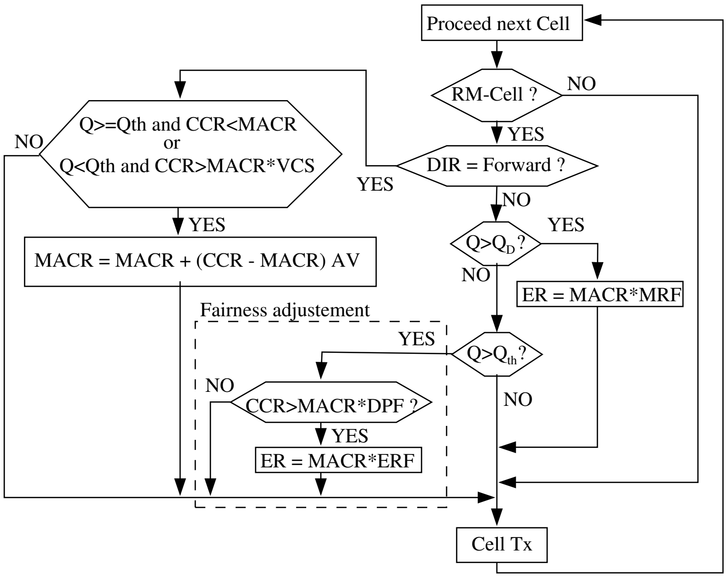


Figure 5.1: EPRCA switch mechanism.

instead of setting the EFCI bit [1].

A main problem of this switch mechanism is its lack of fairness. For example, RM-cells of a VC going through a higher number of congested links will be set to congested more often than those of VCs going through fewer congested links. This undesirable effect (known as the "beat down problem") will result in a lower transmission rate for such VCs.

5.2.2 EPRCA Switch

The Enhanced Proportional Rate Control Algorithm (EPRCA) [68] is an improved version of the original Rate-Control algorithm. The flow chart of figure 5.1 shows the switch algorithm. Here, the switch computes an heuristic approximation of a fair transmission rate, equal to the link capacity minus the capacity of the constrained VCs over the non constrained VCs (max-min criterium). The fair transmission rate (MACR in the figure) is computed during the uncongested periods as an exponential average (MACR=MACR+(CCR-MACR)AV) over all the VCs whose CCR is larger than MACR*VCS. AV is the averaging factor and VCS is a VC Separator used to distinguish between constrained VCs by the switch and otherwise constrained VCs. To avoid the "beat down problem", during congested periods the switch just reduces the ER field of the backward RM-cells with a CCR greater than MACR*DPF. The ER is reduced to MACR*ERF. The Down Pressure Factor (DPF) is used to cause the rate setting control when the ACR reaches a value slightly lower than the MACR. The Explicit Reduction Factor (ERF) is used to set the explicit rates slightly below MACR so that the switch will stay uncongested.

The switch is considered congested when the queue length (Q) is greater than a threshold (Qth). If Q is greater than another threshold Q_D, the switch is considered very congested and ER is reduced in all backward RM-cells to MACR multiplied by MRF (a major reduction factor).

In [68], the following values are suggested: VCS=7/8, DPF=7/8, ERF=15/16, MRF=1/4, AV=1/16. These values are used in the EPRCA switch simulated in this chapter.

[1] The EFCI switch simulated in this chapter makes use of the CI bit of backward RM-cells.

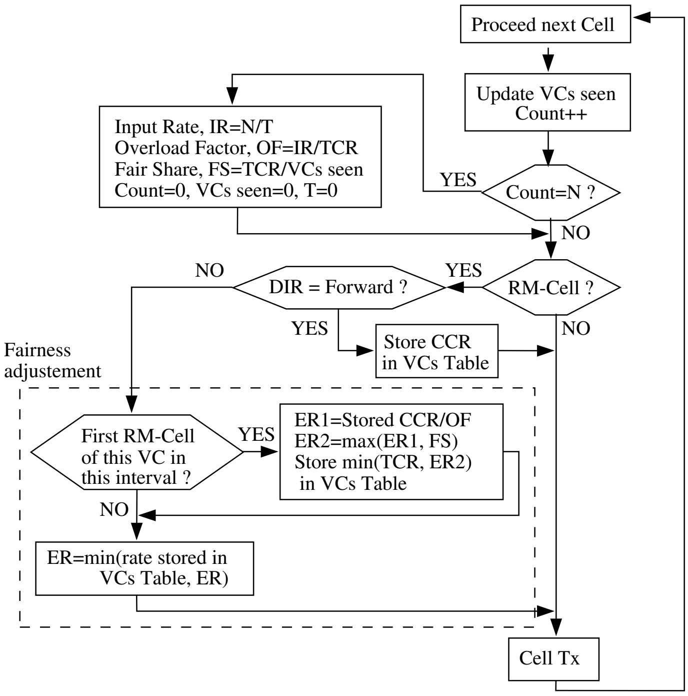


Figure 5.2: ERICA switch mechanism.

5.2.3 ERICA Switch

The Explicit Rate Indication for Congestion Avoidance (ERICA) algorithm [41] is a proposal that tries to keep the queue length low and achieve a max-min fairness. The switch mechanism is described in the flow chart of figure 5.2.

A main difference with the previous switch mechanisms is the detection of the congestion state. In the previous mechanisms this detection is based on a queue length threshold. In the ERICA proposal the switches measure the input rate (IR) and compare it with a target cell rate (TCR, set to 85-95% of the link bandwidth) to compute the overload factor OF=IR/TCR. The ER field of backward RM-cells is then reduced by the OF in order to avoid the congestion state.

To compute the IR, the switch measures the time T until N cells arrive. Then it computes IR=N/T and starts another measuring interval. During each measuring interval, the switch also counts the number of active VCs in order to compute the fair share (FS) as FS=TCR/Number of VCs seen during the measuring interval.

When receiving a backward RM-cell the switch computes the explicit rate ER2 based on load and fairness = max(CCR/OF, FS) and stores the rate min(TCR, ER2) in a VC table. Furthermore, if the ER field of the cell is higher than this value, the field is replaced by the stored rate. Note that storing information relative to the state of each VC in order to compute the ER values is also an important difference with respect to the switch algorithms described in sections 5.2.1 and 5.2.2.

To reduce the feedback delay the algorithm do not use the CCR carried by the backward RM-cell in the ER computation, but the CCR seen in the last forward RM-cell of the same VC. Therefore, this value is also stored in the VC table when a forward RM-cell is received.

To compute the ER the switch uses the IR and FS values computed in the previous interval.

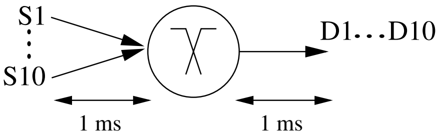


Figure 5.3: Network configuration.

PCR (Cells/ms)	MCR (Cells/ms)	ICR (Cells/ms)	RIF $(1/2^x)$	RDF $(1/2^x)$	Nrm (Cells)
365	2	10	1	1	32

Table 5.1: Source parameters.

EFCI	EPRCA							ERICA	
Qth	Qth	Q_D	VCS	DPF	ERF	MRF	AV	TCR	N
100	100	1000	7/8	7/8	15/16	1/4	1/16	$0.9 \cdot$ LCR	100

Table 5.2: Switch parameters.

In order to avoid oscillations the proposal establishes also that all the backward RM-cells of a given VC seen during the same measuring interval must use the same stored rate. Therefore, the switch stores the computed rate when the first RM-cell is seen during a measurement interval, and keeps an indication that a backward RM-cell of that VC has been seen in the current measurement interval.

Although not discussed here, there exist many extensions and modifications that try to eliminate some of the problems found in the basic ERICA algorithm previously described [40].

5.3 Simulation results

This section presents a simulation analysis of the switching mechanisms previously described. The switches are assumed with output buffering that feed links of identical capacity (LCR) set to 365 cells/ms (≈ 155 Mbps). The congestion control mechanism is applied independently to the buffer located at each output port.

The return path is assumed to be free of data traffic, so the switch will immediately forward the backward RM-cells. Source and switch parameters used in the simulation are listed in tables 5.1 and 5.2 respectively.

5.3.1 Greedy sources

This section investigates the switch behavior when fed by sources that have always cells to send and transmit at the maximum transmission rate permitted by the network (this kind of sources are commonly called “greedy” sources).

Figure 5.3 shows the network topology used in the simulations. It consists of one switch fed by 10 sources sharing the same output link. Note that the end-to-end delay of the sources is 2 ms, equivalent to a distant of 400 km assuming a propagation delay of 5 μs/km.

The figures 5.5, 5.6 and 5.7 show the transmission rate of the sources, the queue length and the link utilization[2] of the congested port of the switch using the EFCI, EPRCA and ERICA switch mechanisms. The sources have a delayed turn-on of 30 ms.

Figures 5.5 and 5.6 show that the queue threshold mechanism used to detect the congestion with the EFCI and EPRCA switches produces an oscillation behavior of the queue length and the transmission rate of the sources. The peak queue length of the EFCI switch is specially sensitive to the number of VCs. The EPRCA switch achieves a better control of the queue length due to the fast decrease of the ACR obtained through the ER feedback.

Figure 5.7 shows that the ERICA mechanism avoids oscillations when the steady state is reached. This is because the ERICA switch algorithm continuously modifies the ER field of the backward RM-cells to the computed value based on fairness, equal to TCR/10. The peak queue length is kept low because the TCR is set to 0.9·LCR. This causes a reduction of the ER based on load before the input rate reaches the LCR, having the effect of a preventive control. Note however that the bandwidth shared by the sources during the steady state is TCR and not LCR. Table 5.3 summarizes the peak queue length and link utilization measured in each simulation.

	EFCI	EPRCA	ERICA
Peak queue length	3172	2408	118
Link utilization (%)	90	92	89

Table 5.3: Peak queue length and link utilization measured multiplexing 10 greedy sources with the EFCI, EPRCA and ERICA switch mechanisms.

5.3.2 Fairness analysis

Figure 5.4 shows the network topology that has been considered to investigate the fairness of the switch mechanisms. Each of the switches of this network has a congested port fed by two greedy sources. The source S1 crosses three congested ports while the sources S2, S3 and S4 cross only one congested port. Note that all the sources have the same end-to-end propagation delay of 2 ms.

Figure 5.8 shows the traces of the transmission rate of the sources using EFCI, EPRCA and ERICA switches. The trace obtained using EFCI switches shows that the source S1 gets a lower transmission rate than the others. This is because the first source crosses three congested ports while the other sources only cross one congested port. The unfairness observed in this trace illustrates the beat down problem explained in section 5.2.1.

The fairness index has been computed using formula (4.5). Since the congested ports are fed by 2 sources, the fair allocation is an equal bandwidth for all the sources. Table 5.4 shows the measured mean transmission rate of each source, the resulting fairness index and the utilization of the congested port. The table reflects the beat down problem of the EFCI switch mentioned before. The EPRCA switch improves the fairness index due to the fairness adjustment of the mechanism. The ERICA algorithm equally distributes the bandwidth achieving a fairness index of 1.

[2] The link utilization is computed averaging the output rate over intervals of 0.4 ms.

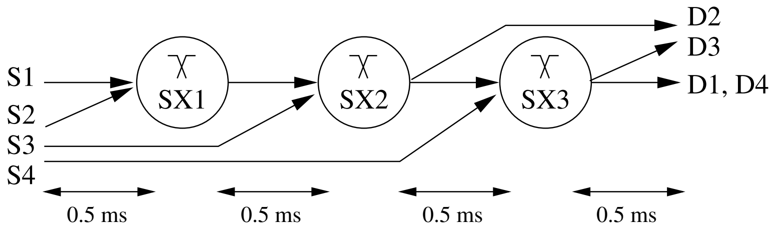


Figure 5.4: Network configuration.

	Mean Tx rate (Cells/ms)				Fairness index	Link utilization (%)		
	S1	S2	S3	S4		SX1	SX2	SX3
EFCI	81	246	218	203	0.897	89	82	78
EPRCA	152	201	194	191	0.989	96	94	93
ERICA	162	162	162	162	1.0	89	89	89

Table 5.4: Mean ACR, switch utilization and fairness index measured multiplexing 4 greedy sources with the EFCI, EPRCA and ERICA switch mechanisms.

5.4 Conclusions

The specification of ABR given by the ATM Forum allows a diversity of switch algorithms. In fact, algorithms for ABR switches have been the subject of an intense research during the last years.

In this chapter three switch algorithms are analyzed, namely the EFCI, EPRCA and ERICA, to show the different degrees of performance and complexity that can be achieved. The analysis is done through simulation, assuming delays of the order of ms.

EFCI is the simplest switch mechanism. It only monitors the queue length and marks the backward RM-cells when higher than a threshold. However, simulation results show that high queue length can be reached and fairness cannot be guarantee.

In order to improve performance, the EPRCA switch mechanism computes an average ACR reading the CCR field of forward RM-cells and modifying the ER field of backward RM-cells. This switch achieves a better performance in terms of link utilization, queue length and fairness than the EFCI switch.

The ERICA switch mechanism introduces the highest degree of complexity. It requires measuring the input rate of each buffer and accessing to a VC table each time a forward or a backward RM-cell is received. It achieves however a high degree of fairness and a tight queue length control. Another advantage in front of the EPRCA is the small number of parameters to be tuned (the target utilization and the measuring interval in cells).

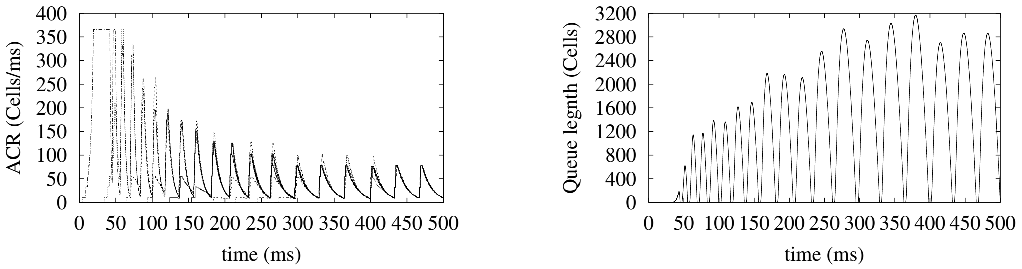


Figure 5.5: Transmission rate of the sources and queue length when multiplexing 10 greedy sources using the EFCI switch mechanism.

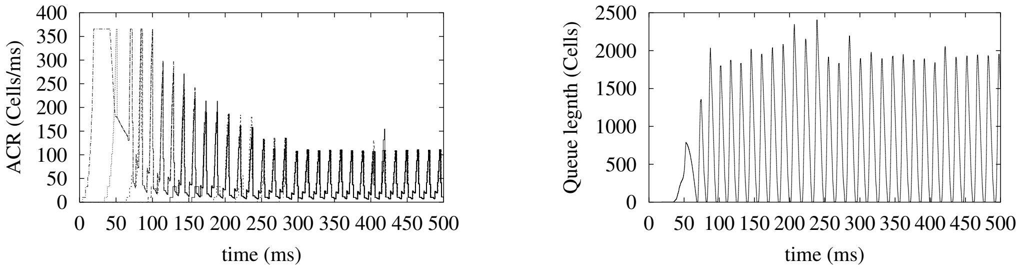


Figure 5.6: Transmission rate of the sources and queue length when multiplexing 10 greedy sources using the EPRCA switch mechanism.

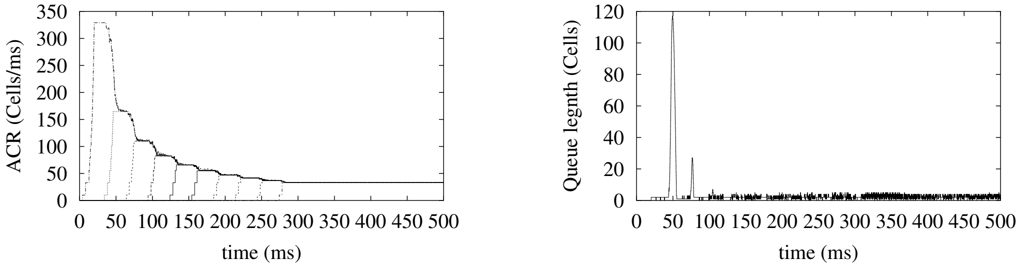


Figure 5.7: Transmission rate of the sources and queue length when multiplexing 10 greedy sources using the ERICA switch mechanism.

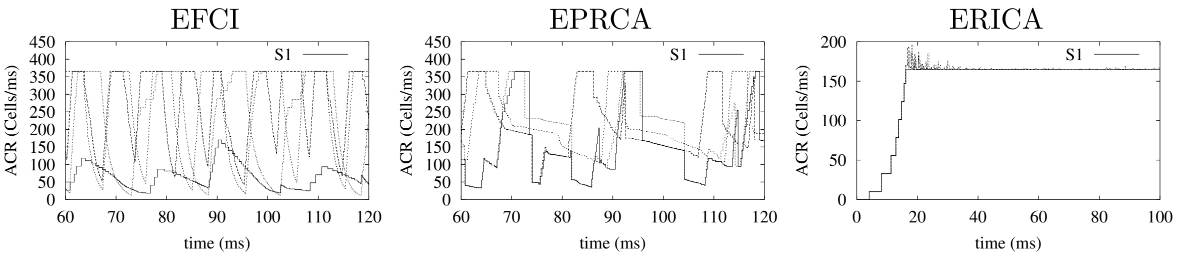


Figure 5.8: Source rates when multiplexing 4 greedy sources using the EFCI, EPRCA and ERICA switch mechanisms with the network topology of figure 5.4.

Chapter 6

Experimental Analysis of an ER Switch

6.1 Introduction

Due to the complexity of the ABR flow control, ABR switches have been mainly evaluated by simulation, as shown in the previous chapter. Analytical studies of the ABR performance can be found considering simple scenarios and applying different simplifications. For example, in [63] ABR is analyzed with a fluid model, while a queuing model is applied in [64, 8].

However, in the research on ABR, one may find a lack of experimental work. The reason may be explained by the delay on the introduction of ABR products on the market. In this chapter one of the pioneer commercial ABR switches developed by ATM Systems [6] is described. The switch performance is analyzed by means of a first set of experiments carried out at the EXPERT Testbed [27]. In the following the goals of the experiments are introduced.

As explained in chapter 4, the main objective of the ABR service is achieving a high network utilization by dividing the bandwidth left by the guaranteed rate traffic among the ABR sources. This chapter focuses on this goal. Furthermore, in this chapter it is called *ABR efficiency* to the network ability to fill the available bandwidth with the ABR traffic. A numerical value of this measure is computed by means of the following relation:

$$\text{ABR efficiency (\%)} = \frac{\text{Bandwidth filled by the ABR traffic}}{\text{Available bandwidth}}\,\text{x}\,100 \tag{6.1}$$

In the chapter the effect of a long distance delay on the feedback loop is also investigated.

The rest of the chapter is organized as follows. In section 6.2 the ABR equipment used in the experiments is described. Section 6.3 gives some details of the measuring framework and the traces that were obtained. Finally, section 6.4 gives some concluding remarks.

6.2 Description of the ABR Equipment

An ABR SES/DES source and an ABR switch kindly provided by *Able Communications* [1] and *ATM Systems* [6] respectively was available at the testbed. The ABR SES/DES source was the

AC-1000 ATM Protocol Analyzer and the switch under test was the 8100 with a 16 ports card STM1/OC3c. In the following a brief description of the switch algorithm is given.

The exact switch algorithm was not known, but the following guidelines were provided by ATM Systems. The switch counts the number of active VCs and traffic over all of them at each measuring interval given by a small period of time. Then the switch computes the so called Maximum Allowable Cell Rate (MACR) as:

$$\mathrm{MACR} = \frac{\text{bandwidth available to locally bottlenecked VCs}}{\text{number of locally bottlenecked VCs}} \tag{6.2}$$

The bandwidth available to locally bottlenecked VCs is the total bandwidth used by locally bottlenecked VCs in the last measuring period plus all bandwidth used by UBR and all unused bandwidth. VCs that are near to or exceed the MACR are considered locally bottlenecked.

The ER field of backward RM-cells is reduced down, if higher, to the MACR times a congestion correction factor. This factor is near one for low congestion and typically 50% if a certain queue limit is exceeded and the queue is still growing. In order to speed up the feedback loop, the RM-cells have priority over the ABR data cells.

In order to fulfill the bandwidth guarantees the switch implements a Weighted Fair Queuing scheduling algorithm which works as follows. The VCs are each assigned to one of 16 queues. Each queue has similar traffic, say all ABR or all CBR. There is a Rate Guarantee computed for each queue which is the sum of all the VC rate guarantees for the queue plus some additional bandwidth if desired. Each queue has a Reset Counter value given by the Link Cell Rate divided by the Rate Guarantee. At each cell time slot the switch counts down each queue counter by one and compares all the 16 counters. The cell of the queue with the lowest counter is sent, and the counter set to the Reset Counter value.

6.3 Performance Results

The performance of the ABR switch was analyzed in two scenarios regarding the involved delays. In the first one all the interconnections where inside the testbed with a cable length of several meters, therefore the propagation delays could be neglected. This scenario is referred to as the *LAN scenario*. In the second scenario a connection of the Pan European ATM Network was used to investigate the influence of a long distance delay. Thus, it is referred to as the *WAN scenario*. In both scenarios a single VBR source was multiplexed with a single ABR greedy source using the topology shown in figure 6.1. In the following the equipment, interconnections, and parameters used to obtain the traces are first described, and then, the traces obtained are explained.

6.3.1 Description of the Testbed Configuration

In order to set up the topology of figure 6.1, the physical configuration shown in figure 6.2 was used. Note that the bottlenecked link in figure 6.2 corresponds to the output link of port 4 of the *ATM Systems* switch. The Pan European connection used in the WAN scenario was a loopback from the testbed in Basel to Oslo and introduced a delay of 38.25 ms. The delay was introduced in the return path of the backward RM-cells, between the congested port and the SES (port A1 of the *FORE* switch). Note that the feedback control depends only on the

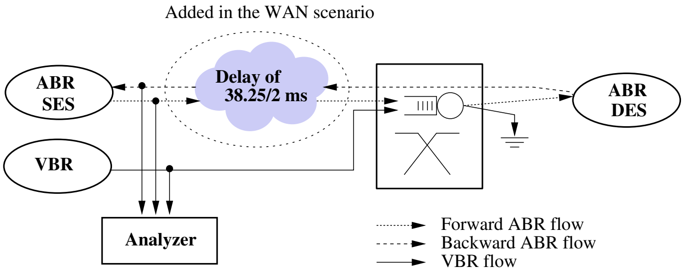


Figure 6.1: Experiments logical configuration.

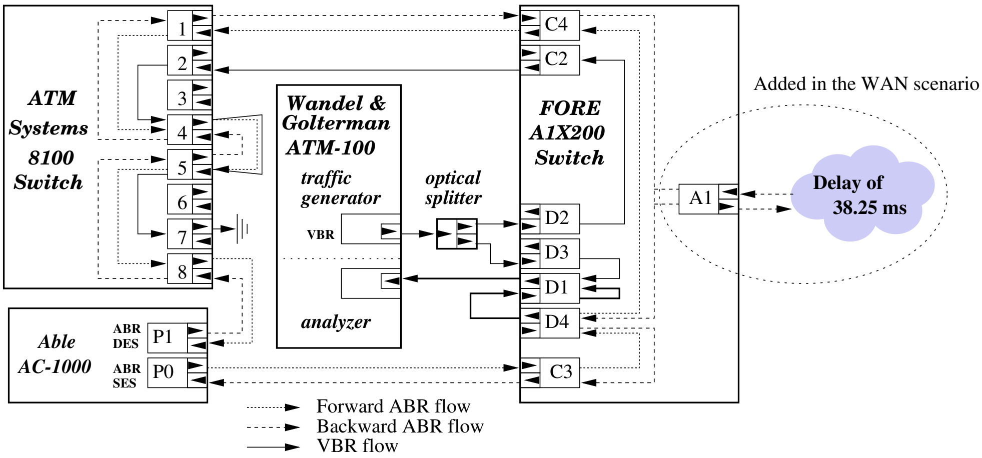


Figure 6.2: Experiments physical configuration.

overall round trip delay between the switch and the SES. Therefore, the configuration would be equivalent to introducing a delay of 38.25/2 ms in the forward and backward streams, as indicated in figure 6.1.

The *ATM-100* analyzer is able to record a trace in an ASCII file made up of a time stamp of each cell arrival, the decoded header and the payload contents. There is only one port available to record a trace, so all the streams to be measured were multiplexed on the output link of the port D1 of the *FORE* switch, and the output of this port was fed into the *ATM-100* analyzer. 64000 cells where recorded for each trace. By means of the VPI-VCI values, the different streams could be isolated from the common trace. This multiplexing stage introduced a jitter on the individuals cell streams to be measured. However, this jitter is negligible since the output link of the *FORE* switch is an STM1 (155.52 Mbps), a rate much higher than the measured aggregate cell stream.

Note that an optical splitter was used to separate a copy of the VBR stream for measuring. To get a copy of the forward and backward ABR streams, a multiplexing stage of the *FORE* switch (port D4) was used instead. This was done that way because only one optical splitter was available. The *FORE* switch was also used to convert the optical multiple mode fiber connections of the *ATM-Systems* ports to the single mode ones of the *ATM-100*.

The ABR source parameters adjusted in the *Able* box are given in table 6.1. Note that RIF is set to 1, therefore the ABR source modified the rate to the ER field of backward RM-cells. Although the *ATM-Systems* ports were STM1/OC3c cards of 155.52 Mbps, at the moment of the experiments a patch had to be introduced in the switch which limited the capacity to 2 Mbps

PCR (Cells/s)	MCR (Cells/s)	ICR (Cells/s)	RIF ($1/2^x$)	Nrm (Cells)	Trm (ms)
4717	0	4717	0	32	100

Table 6.1: ABR source parameters.

(4717 Cell/s). Finally, the measuring interval of the *ATM-Systems* switch was configured to 4 ms.

6.3.2 Description of the Measured Traces

The impact of available bandwidth pattern on the ABR efficiency was investigated using the LAN scenario. In this case a periodic ON-OFF VBR source with an ON and OFF cycles of equal duration was used. The measured traces are shown in figures 6.3 to 6.8. The figures depict the rates and an estimation of the queue length obtained after processing the traces. The VBR, ABR and VBR+ABR rates are plotted superposing the ER of backward RM-cells with points. Note that the points are plotted at the backward RM-Cell arrivals to the SES. The graphs are plotted at the time interval where the peak queue length was achieved. For sake of clarity the time axis has been shifted such that the traces begin at 0 ms. Table 6.2 summarizes the VBR semi-cycle and PCR used for each trace, and gives the average rates, efficiency and peak queue lengths measured from each trace.

The figures evidence the fact that the ABR source can follow the VBR rate changes as long as the rate changes do not occur at time intervals of shorter duration than the time intervals between backward RM-cells arrivals. Let's have a look for example at figure 6.6. When the VBR source is in the ON state, the ABR source rate is reduced to $\approx$ 255 cells/s, and thus, the time interval between RM-cells is Nrm/255 = 125 ms. This time interval, however, is upper bounded by the Trm source parameter to 100 ms. The figure shows that with such a low backward RM-Cell rate, the gaps left by the VBR source cannot be cached up by the ABR source, and only an efficiency of 13 % is achieved (cfr. table 6.2). Figure 6.3 shows that although the semi-cycle is 5 times smaller than the previous one, the RM-Cell rate is high enough to approximately follow the VBR variations, resulting in an efficiency of 70 %. Therefore, achieving a high ABR efficiency in a fast varying available bandwidth pattern would require reserving a certain amount of bandwidth for the ABR traffic. This consideration could be a CAC concern.

In the WAN scenario, only the trace shown in figure 6.9 was taken. The figure shows that the consequence of the delay is the longer reaction time of the ABR source. The VBR source becomes active at 150 ms. The switch detects the available bandwidth reduction in the following measuring interval (remember that this interval is set to 4 ms). This reduction is applied to the next backward RM-Cell going through. The rate reduction conveyed by this backward RM-Cell, however, takes place at the switch after the round trip feedback delay of 38.25 ms. A queue length up to 196 cells is built up due to the overload held up during the overall feedback delay. Remember that in the experiments a link cell rate of 2 Mbps was used. Taking into account that the queue length is proportional to the transmission rate, an equivalent experiment with a 155 Mbps link would have yield a peak queue length of $196 \cdot 155/2 = 15190$ cells approximatively. The conclusion is that in a WAN environment the ABR service is expected to have long queue lengths (that may be in the order of thousand of cells).

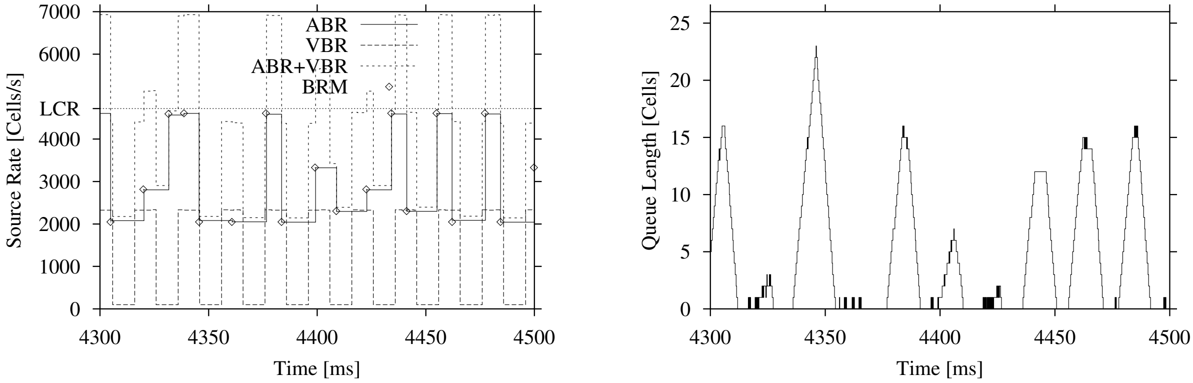


Figure 6.3: LAN scenario. VBR semicycle = 10 ms, ON-Rate = 1 Mbps (2358 cells/s).

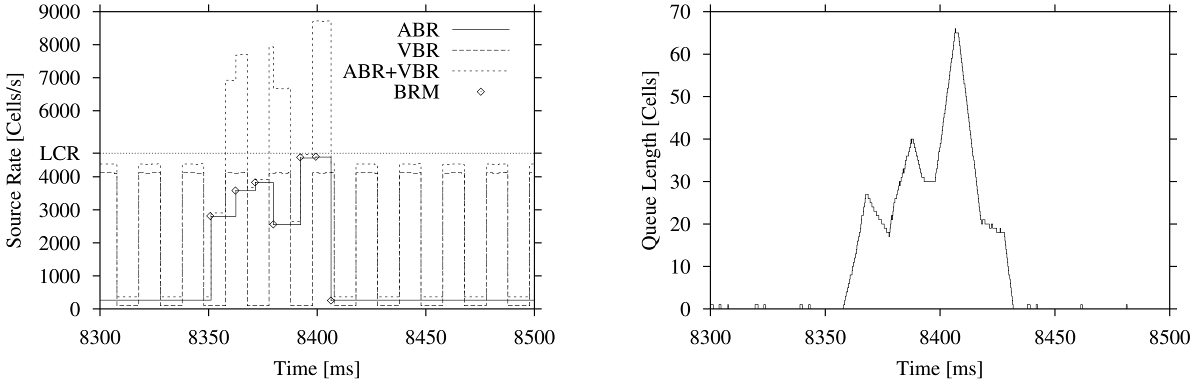


Figure 6.4: LAN scenario. VBR semicycle = 10 ms, ON-Rate = 1.8 Mbps (4245 cells/s).

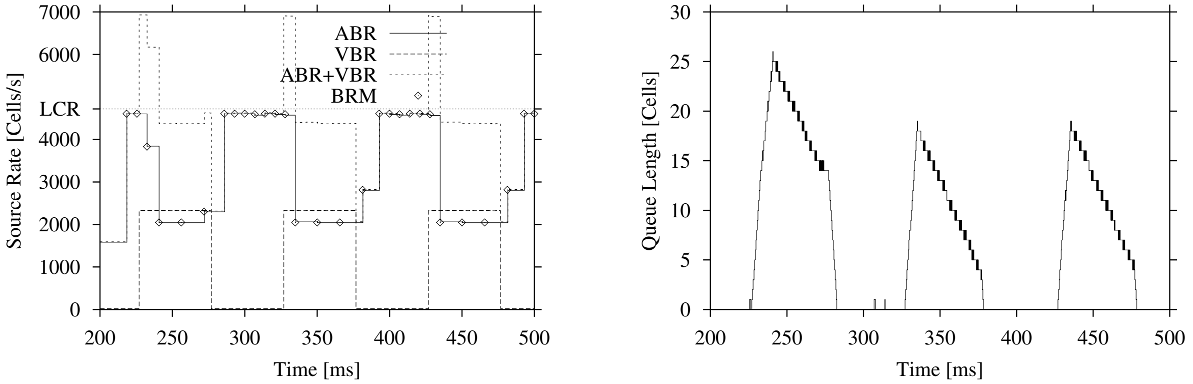


Figure 6.5: LAN scenario. VBR semicycle = 50 ms, ON-Rate = 1 Mbps (2358 cells/s).

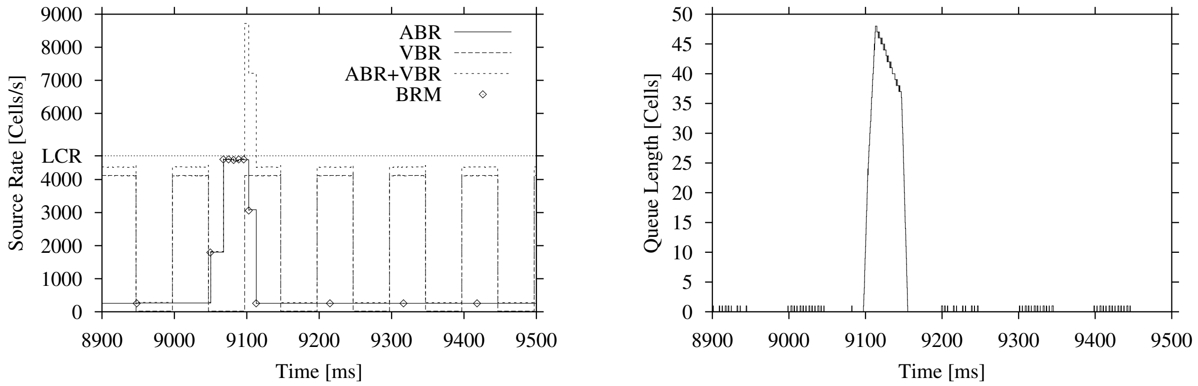


Figure 6.6: LAN scenario. VBR semicycle = 50 ms, ON-Rate = 1.8 Mbps (4245 cells/s).

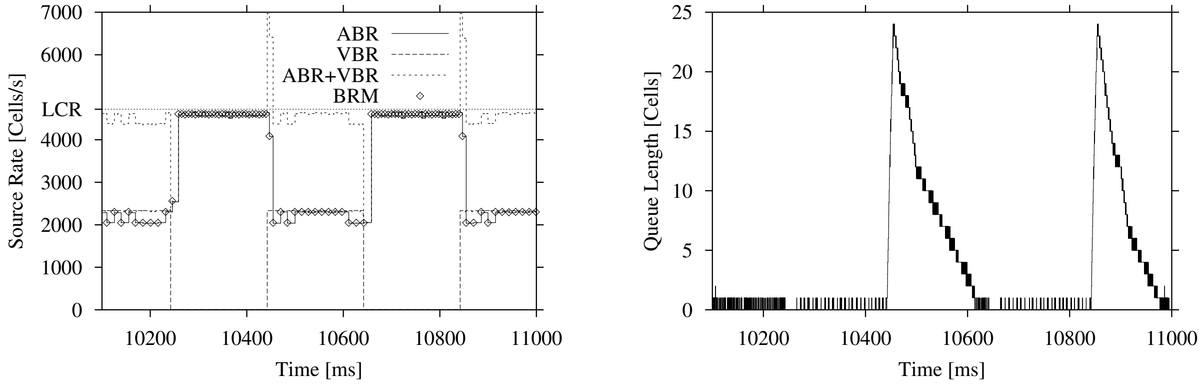


Figure 6.7: LAN scenario. VBR semicycle = 200 ms, ON-Rate = 1 Mbps (2358 cells/s).

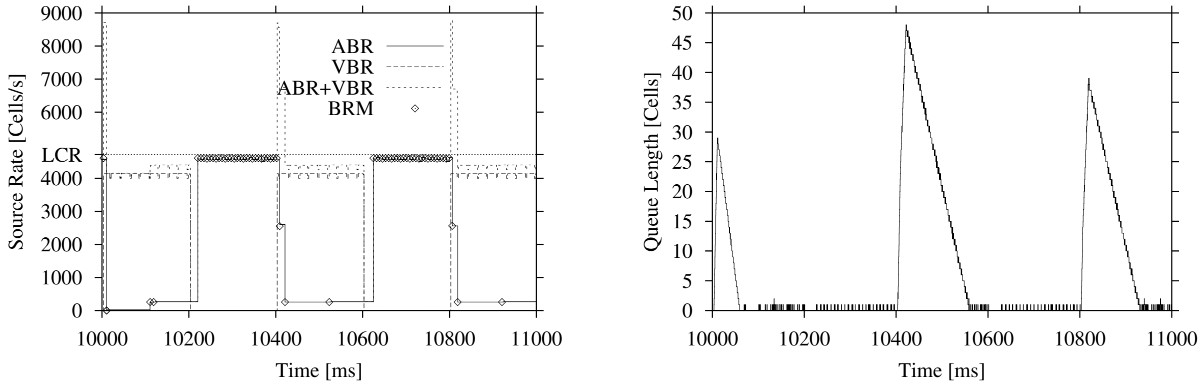


Figure 6.8: LAN scenario. VBR semicycle = 200 ms, ON-Rate = 1.8 Mbps (4245 cells/s).

VBR Semicycle (ms)	VBR PCR (Cells/s)	VBR av. rate (Cells/s)	Avail. Bw. (Cells/s)	ABR av. rate (Cells/s)	Efficiency (%)	Peak Q.L. (Cells)
10	2358	1200	3517	2487	70	23
	4245	2100	2617	591	22	66
50	2358	1170	3547	2787	78	26
	4245	2064	2653	352	13	48
200	2358	1167	3550	3162	89	24
	4245	2061	2656	2309	86	48

Table 6.2: Summary of figures 6.3 to 6.8.

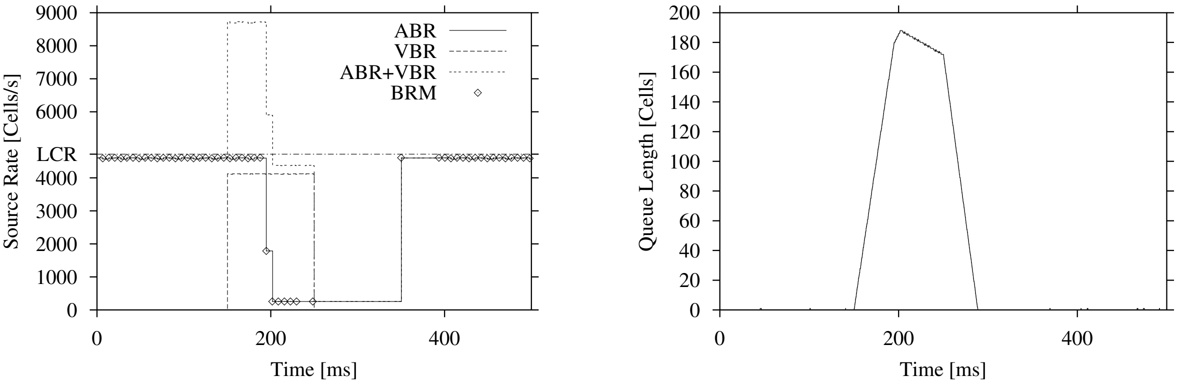


Figure 6.9: WAN scenario. VBR ON-Rate = 1.8 Mbps (4245 cells/s).

6.4 Conclusions

ABR has been extensively analyzed during the recent years, but there is a lack of practical experiments on ABR due to the delay on the introduction of ABR products on the market. In this chapter a first set of experiments carried out at the EXPERT Testbed [27] with a pioneer ABR switch developed by ATM Systems [6] is presented.

The chapter focused on the ABR efficiency when the network resources are shared with guaranteed rate traffic, and the influence of a long distance delay. These issues have been studied multiplexing an ABR with a VBR source. In the chapter the equipment, configuration and method to obtain the traces is described. Furthermore, rates and queue lengths obtained from the processed traces are depicted.

The figures show that the ABR efficiency may be very sensitive to the VBR traffic pattern. The figures evidence the fact that in order for the ABR source to be able to follow the VBR rate changes, a certain minimum RM-cells rate is needed. This consideration could be a CAC concern. The measures obtained with the long distance delay show that in a WAN environment the ABR service is expected to have long queue lengths.

Chapter 7

Improvements of ER Switch Algorithms

7.1 Introduction

Remember from chapter 5 that ER switches reduce down the ER field of the RM-cells in order to convey the fair rate to the sources. When receiving a backward RM-cell the sources adjust the transmission rate at most to the conveyed ER. However, if a source becomes idle or internally rate limited it may transmit at a lower rate than the offered ER. Therefore, the bandwidth left by such sources should be redistributed among the others in order to achieve a high utilization.

The Current Cell Rate (CCR) field of the RM-cells has been defined to inform the network about the rate at which the sources are actually transmitting (see chapter 4). This field is set to the Allowed Cell Rate (ACR). Therefore, in order the CCR to be useful the ACR should be maintained close to the transmission rate of the source. The ATM-Forum, however, has not specified any rule to constrain the sources to do that. Moreover, the Conformance Definition proposed for ABR (the DGCRA) does not perform a CCR test, so the network cannot trust the CCR. A detailed study of the DGCRA is given in chapters 8 and 9.

In this section modifications in order to solve these drawbacks are proposed. To illustrate some of the benefits of this proposal we also explain how the popular ERICA switch could be modified to measure the ABR input rate and number of active VCs based on the conveyed CCR. The ERICA algorithm is simplified and similar modifications could be applied to other switch mechanisms, i.e. [3, 43, 77].

7.2 Motivation of the Proposal

7.2.1 Importance of the CCR

As explained in section 4.4.1, the Max-Min fairness criteria has been mainly used as a goal for ABR switch mechanisms. Remember that this algorithm consists of allocating to the unconstrained sources a rate given by:

$$B = \frac{A - U}{N - N'} \tag{7.1}$$

where A is the available bandwidth on the link, N the total number of active sources sharing the link, U the sum of bandwidth of the constrained sources and N' its number.

For example, the switch algorithm described in [77] computes the fair share applying the formula (7.1). In this framework the switches modify the ER of both forward and backward RM-cells. By keeping track of the conveyed ER of forward and backward RM-cells the switches determine whether the sources are constrained at the local link or at other switches.

An algorithm exclusively based on the conveyed ER, as the one previously described, is only efficient if sources use all the bandwidth conveyed by the ER field. If there are sources which become silent or internally rate limited the bandwidth offered to them would be wasted. It is therefore important that switches check the transmission rate of the sources to redistribute the unused bandwidth.

To communicate to the network the transmission rate the ATM-Forum establishes that the sources have to set the CCR field of forward RM-cells to the ACR. Remember that the ACR is a source parameter which fixes the maximum rate at which cells may be scheduled for transmission. This parameter is assumed to be close to the transmission rate of the source.

The switch algorithm proposed in [43], for example, is based on formula (7.1) but takes into account the CCR to determine if sources are congested at the local link or elsewhere.

Another switch algorithm which takes into account the transmission rate of the sources is the popular ERICA explained in section 5.2.3. Remember that this algorithm does not directly apply formula (7.1). Instead, the switch measures the ABR input rate in order to compute the overload factor z = ABR input rate / ABR Capacity, and the Fairshare = ABR Capacity / Number of active ABR Sources. Source rates conveyed by the CCR field of forward RM-cells are stored in a table. When receiving a backward RM-cell the switch uses the CCR to compute the VCShare = stored CCR / z. Then, the ER field of the backward RM-cell is reduced, if greater, down to max(Fairshare, VCShare).

7.2.2 The problems of the CCR in the current ATM-Forum Specification

With the current ATM Forum specification the CCR may not be an accurate measure of the source rates because:

1. The DGCRA proposed by now as the policing function does not perform a CCR conformance (i.e. does not check that the sources effectively change the CCR with the ACR), so trusting the CCR may lead to a misbehavior of the feedback control if any source sets this value erroneously.

2. The ATM-Forum has not specified any rule to constrain the sources to maintain the ACR close to the transmission rate. Thus the CCR could yield an overestimated value.

3. In the ATM-Forum specification a source that becomes silent stops sending RM-cells. Checking the CCR of RM-cells is therefore not enough to keep track of the source transmission rate.

To avoid these drawbacks the ERICA switch does not use the conveyed value by the CCR to compute the input rate and number of active VCs. These are computed each "Measuring Interval" given by the N cell arrivals. The input rate is computed as $N/$time interval, and the number of active VCs by counting the VCs with at least one cell arrival during the measuring interval. These solutions however, do not completely solve the problem because the algorithm uses the CCR when computing the VCShare. To completely avoid using the CCR a Per-VC CCR Measurement Option is proposed in [40]. This option consists of measuring the per VC rates by taking ratios of cell counting and time intervals in a similar way the ABR input rate is measured. This option however would increase the complexity of the switch a lot. The following sections propose solutions to the drawbacks of using the CCR.

7.3 A UPC Based on the CCR

In this section a UPC based on the Current Cell Rate (CCR) is proposed. A further description of this proposal is given in section 8.3.

The Sources set the CCR of forward RM-cells to the ACR. The idea is thus using the CCR conveyed by the forward RM-cells to perform the rate conformance. The proposed UPC based on the CCR would perform a conformance test at two levels. At a first level it would check that cells are conforming to the rate conveyed by the CCR, and at the second level it would check that the source correctly changes the ACR by means of the expected rate at the measuring point. Such a UPC would guarantee that the source rate will never exceed the value conveyed by the CCR, and thus it could be safely used by the switches. A CCR non-conformance condition would occur in case the UPC receives a forward RM-cell with the CCR exceeding the expected rate. In this case several actions could be taken by the UPC which would be of standardization latitude, e.g. the CCR could be reduced down to the expected rate or the RM-cell could be simply discarded.

With the current source behavior given by the ATM Forum a source may transmit cells following a forward RM-cell at a rate up to the CCR until a backward RM-cell conveying a different rate is received. Consequently, with such source behavior, it is not possible to use the CCR conveyed by the forward RM-cells to perform the rate conformance because at any time cells may be received at a higher rate. We propose to modify the source behavior constraining it to delay the rate increases just after a forward RM-cell transmission.

However, remember that the RM-cells are transmitted after each (Nrm-1)th data-cell. Therefore, if the source ACR is low the time interval between RM-cells could be too large, slowing down the ramp-up of the sources and consequently the access to the unused network bandwidth. To solve this drawback we propose allowing the sources to advance an RM-cell transmission before its Nrm-th turn when receiving what we call a "worth" rate increase. Advancing RM-cell transmission would increase the RM- cell rate and thus the overhead introduced by the RM-cell transmission. We therefore introduce the concept of an "RM-cell advance algorithm" which would decide whether rate increases are worth to advance an RM-cell transmission or not.

A simple RM-cell advance algorithm would be the following. Let PACR be the rate increase conveyed by a backward RM-cell. The algorithm would consist of advancing a forward RM-cell transmission if PACR $\geq \alpha \cdot$ ACR, $\alpha > 1$. Clearly, the higher the value of α, the higher the rate increase required to perform the RM-cell advance, and thus the less times such advance will be performed. Note that setting $\alpha = 1$ would imply performing the RM-cell advance at any rate increase.

7.4 An Accurate Rate Indication by Means of the CCR

The CCR field of RM-cells conveys the ACR parameter of the sources. Therefore, in order the switches to accurately measure the source rate by means of the CCR, sources should be encouraged to maintain the ACR close to the actual transmission rate, e.g. implementing the optional use-it-or-lose-it behavior. To achieve this goal we propose charging the connection based, at least in part, on any measure of the conveyed CCR (e.g. the time average). Note that in the ABR framework the network offers a fair rate to the sources by means of the ER (which may exceed their needs). Then the sources set their rate at most to the received ER and communicate the rate to the network by means of the CCR. Therefore, pricing the CCR is a monetary incentive for the sources to dynamically equilibrate their rate demand and the network capacity and thus improve the network performance.

In the framework described in section 7.5 the sources delay rate increases just after a forward RM-cell transmission but perform rate decreases immediately. Consequently, if the source rate sharply drops, because of a rate change conveyed by a backward RM-cell or because the source became internally rate limited, the source maybe interested in advancing a forward RM-cell transmission to signal the rate decrease. Moreover ON-OFF sources should indicate the silent periods.

The ATM Forum standard does not establish any indication mechanism for the silent periods. Knowing the number of active sources is however a required parameter of many switch mechanisms in order to apply the fairness algorithm [3, 43, 77, 40]. In many cases no mechanism is described to compute the number of active sources [43, 77] or it is assumed to be the number of established connections [3]. With this assumption however, the unused bandwidth of idle connections would be wasted. The ERICA switch [40] computes the number of active sources by counting the VCs with at least one cell arrival during the measuring interval. This mechanism is however a burden for the switch which requires to check each cell arrival. Moreover, if there are sources transmitting at very different rate, or if the measuring interval is not large enough, it may not properly measure the number of active sources.

We propose delimiting activity periods by two RM-cells as described in the following. The sources would start transmitting an ON period as establish by the ATM Forum after the connection setup, i.e. setting the ACR at most to the ICR and sending a forward RM-cell. After an ON period sources would transmit a trailer RM-cell conveying a CCR equal to the MCR. In case of a MCR equal to zero, this trailer RM-cell would reset the UPC to the initialization conditions and advertise switches to decrement the counter of active VC.

Finally, note that in case a forward RM-cell is lost, it does not have a serious influence in the rate control because UPC and switches are initialized at each RM-cell arrival. Just in case a trailer RM-cell is lost, switches would not reset the allocated bandwidth. We thus propose that sources ensure that trailer RM-cell are not lost. This could be done, for example, by sending periodically trailer RM-cells until the corresponding backward RM-cell is received.

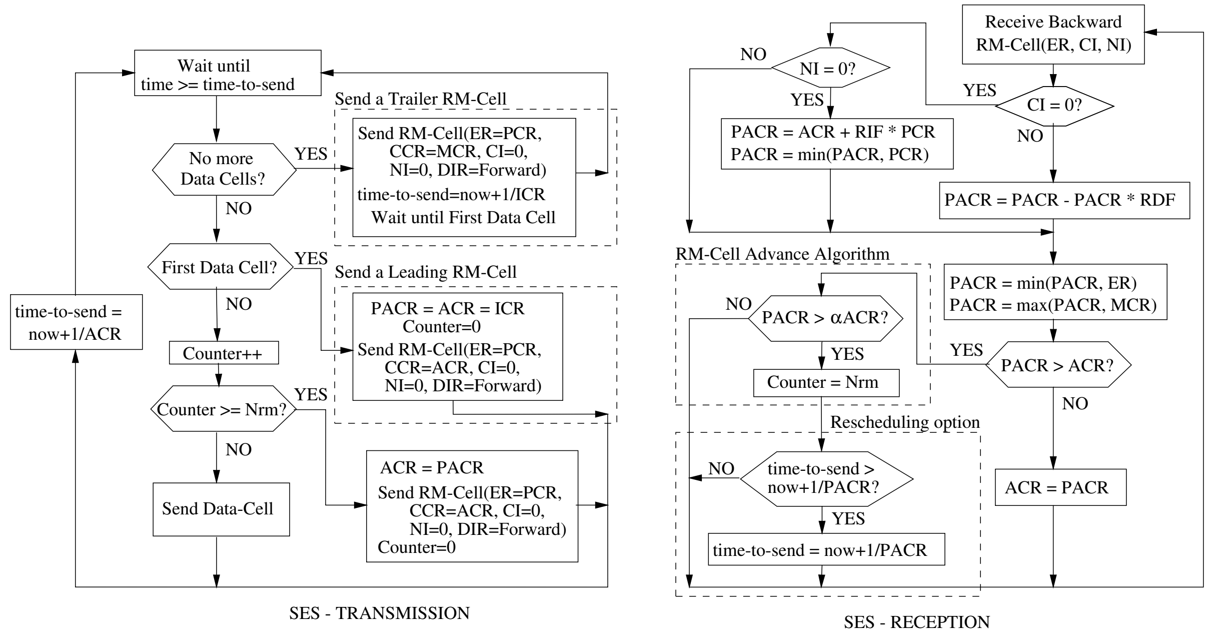


Figure 7.1: Modified source behavior.

7.5 Implementation Example

Modifications on the source behavior proposed in previous sections can be summarized in the following points:

- If a backward RM-cell is received conveying a rate decrease it takes place immediately. If a rate increase is conveyed, it is delayed until next forward RM-cell transmission.
- The source is allowed to send a forward RM-cell before its Nrm-th turn in order to take advantage of a rate increase or to signal a rate decrease. What we call *RM-cell advance algorithm* is the responsible of deciding when an RM-cell transmission is advanced. This algorithm would be implementation specific.
- When a source becomes idle it will have to transmit an RM-cell conveying a CCR equal to the MCR. After an idle period the source will set the ACR at most to the ICR and transmit an RM-cell.

Figure 7.1 shows a simplified flow chart of the source behavior given by the ATM Forum (confront with figure 4.4), including the modifications previously described. The auxiliary variable PACR is introduced in order to delay rate increases after an RM-cell transmission. Note from the SES RECEPTION that PACR carries out all rate changes. If a backward RM-cell conveys a rate decrease the ACR is set to the PACR, but if it results in a rate increase the ACR is not changed. Note also from the SES TRANSMISSION that previously to an RM-cell transmission, the source sets ACR = PACR. Therefore, rate increases will be delayed just after a forward RM-cell transmission.

In the SES RECEPTION the RM-cell advance is performed when PACR $> \alpha \cdot$ ACR. When this condition is accomplished Counter is set to Nrm which implies that the next cell transmission will be an RM-cell. After an RM-cell advance the rescheduling option can also be performed.

SES TRANSMISSION shows that at the beginning of a burst, the ACR is set to ICR and a leading RM-cell is transmitted. After the last cell of the burst, a trailer RM-cell is transmitted with the CCR set to the MCR. In case of a MCR equal to zero this RM-cell will reset the UPC and will decrease the switch counter of active VCs.

7.6 ERICA Switch Based on the CCR

In this section we propose modifications of the ERICA switch mechanism in order to take advantage of the source behavior proposed in section 7.5. The modifications consist of measuring the input rate and the number of active VCs by means of the conveyed CCR. The same algorithm could be performed by other switch mechanisms.

Figure 7.2 shows the algorithm. The switch maintains a table with the last conveyed CCR of each source (`stored_CCR[]` in the figure). At each Forward RM-cell arrival the switch determines if a source starts an activity/silent period by checking if the CCR is equal to zero, and thus, easily maintains the number of active VCs (`n_active_VC` in the figure). The input rate (`input_rate` in the figure) is computed by adding the rate increment of each VC (sentence 6).

Measuring of the input rate and the number of active VCs as previously described introduce important advantages over the original ERICA switch mechanism. Now, no measuring interval needs to be defined, eliminating the setting of parameter N of the switch. Note that in the original ERICA algorithm the parameter N needs to be set carefully. Shorter intervals may not properly measure the number of active VCs and the system may become unstable. Longer intervals slow down the feedback increasing then queue lengths. Moreover, in the original ERICA switch the number of active VCs is computed as the number of distinct VCs seen during the last measurement interval. In order to obtain this measure the switch maintains a bit in a table indicating if any cell of a VC has been seen during the measuring interval. This requires accessing the table at each cell arrival and resetting the table at the end of each measuring interval. Now, just the RM-cells need to be checked.

```
   at each Forward RM-cell arrival(CCR, VCI)
   {
1:          if(CCR != stored_CCR[VCI]) {
2:                            // computes the number of active VC
3:                 if(CCR == 0) --n_active_VC ;
4:                 else if(stored_CCR[VCI] == 0) ++n_active_VC ;
5:                            // computes the input rate
6:                 input_rate += CCR - stored_CCR[VCI] ;
7:                 stored_CCR[VCI] = CCR ;
           }
   }
```

Figure 7.2: Measure of the input rate and number of active VCs.

7.7 Performance Results

In this section the performance benefits of the source behavior proposed in section 7.5 are shown. To show these benefits the performance achieved with the ERICA switch is confronted with the

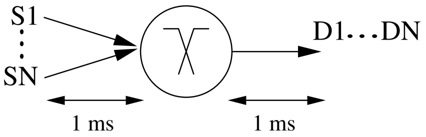


Figure 7.3: Simulation model.

Sources					ERICA Switch		
MCR	PCR	ICR	RIF	Nrm	LCR	TCR	N
0 Cells/ms	365 Cells/ms	2 Cells/ms	1	32	365 Cells/ms	0.9 LCR	100

Table 7.1: Simulation parameters.

modified ERICA switch proposed in section 7.6.

Figure 7.3 shows the network configuration used in the simulations. The sources feed a single switch and share a link of capacity LCR = 365 cells/ms (≈155 Mbps). Sources and switch parameters are given in table 7.1. The RIF of the sources has been set to 1 in order to adjust their rate to the conveyed value by the ER field of backward RM-cells. Two types of sources are considered: greedy sources which have always cells to send and transmit at the maximum permitted rate (the ACR), and ON-OFF sources. For the ON-OFF sources an Open Loop model has been assumed. This consists of the sources having a pool of cells which is filled with bursts of fixed length B. Bursts arrive every ta time period, and the cells are transmitted with a greedy behavior.

We first investigate the time evolution of the ACR of the sources and the queue length of the switch. The sources considered in this simulation are two greedy sources and one ON-OFF source which transmits bursts of $B = 2500$ cells. Bursts arrive at the source deterministically every $ta = 50$ ms. The greedy sources start transmission at time = 0 ms and the ON-OFF source starts the first ON period at time = 20 ms. Figure 7.4 and figure 7.5 show the switch queue length and the ACR of the ON-OFF and one of the greedy sources.

Figure 7.4 depicts the values obtained when the sources behave as described by ATM Forum specification. The ADTF value has been set to 10 ms, so that when the ON-OFF source starts an ON period it will be constrained to reduce its rate to the ICR value. In order to distinguish the ON and OFF periods, the figures plot the ACR value of the source when is ON and 0 when is OFF. When the ON period starts the first cells are transmitted at the previous ACR until the first RM-cell has to be sent. At this moment the ADTF adjustment is performed and the ACR is reduced down to the ICR. This causes the spike which appears at the end of each OFF period.

Figure 7.4 shows two remarkable effects: the oscillations on the ACR of the sources and the picks on the queue length. To explain these effects remember that the ERICA switch computes the ER of RM cells as:

$$\text{ER} = \max(\text{stored CCR / overload factor, fair share}) \tag{7.2}$$

The inaccuracy on the measure of the input rate produces the oscillations on the ACR when the first term of the formula is applied.

The picks on the queue length are produced because when the ON-OFF source becomes active

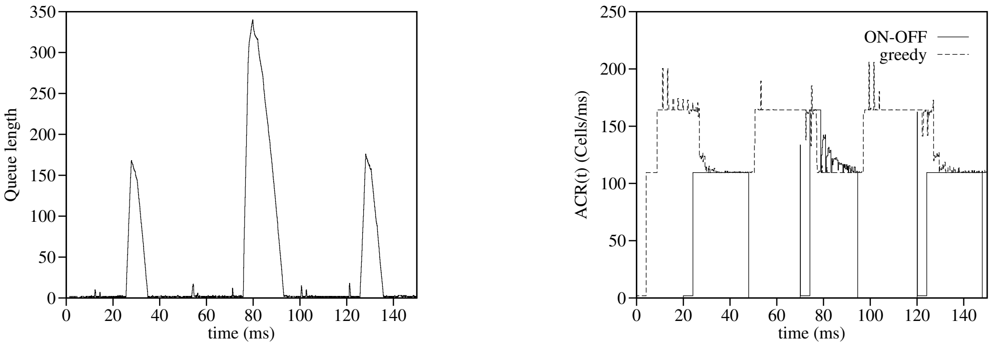


Figure 7.4: ATM-Forum source behavior and original ERICA switch.

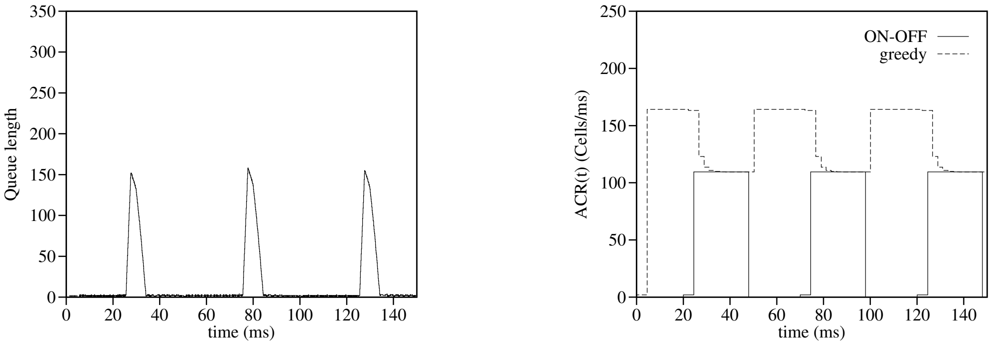


Figure 7.5: Source and ERICA switch behavior proposed in this chapter.

formula (7.2) gives to it the fair share, but the greedy sources maintain their rate because the first term of formula (7.2) yields the maximum. When the cells of the ON-OFF source arrive with the new rate at the switch, the overload factor increases and the ER of the RM-cells belonging to the greedy sources is reduced. When this happens however, the switch is not able to drain the transient overload and a pick queue length is build up.

In the second ON period, figure 7.4 shows that the pick queue length is higher than in the other ones. This is because the ERICA switch measures the number of active VCs by counting the VCs with at least one cell arrival during the measuring interval. When the ON-OFF source becomes active, it starts transmitting at the ICR = 2 cells/ms which is much lower than the transmission rate of the greedy sources ($\approx$ 164 cells/ms). Therefore, in the measuring interval previous to the ER of the ON-OFF RM cell which conveys the first rate increase, just the cells of the other two greedy sources are seen and thus only 2 active VCs are counted. This underestimate of the number of active VCs produces an erroneously computation of the fair share, and the consequent increase on the queue length.

Figure 7.5 shows the switch queue length and the ACR when the sources behave following our proposal described in section 7.5. In this figure, the ERICA switch measures the input rate and number of active VCs perfectly by means of the conveyed CCR, as described in section 7.6. This eliminates the oscillations of the ACR and the underestimate of the number of active VC.

The queue length distribution when the switch is fed by multiple ON-OFF sources is now investigated. 10 ON-OFF sources are considered with parameters equal to the previous simulations but with an interarrival time of bursts exponentially distributed.

Figure 7.6 confronts the complementary cumulative distribution of the queue length when the sources transmit following the ATM Forum specification and the switch is the original ERICA

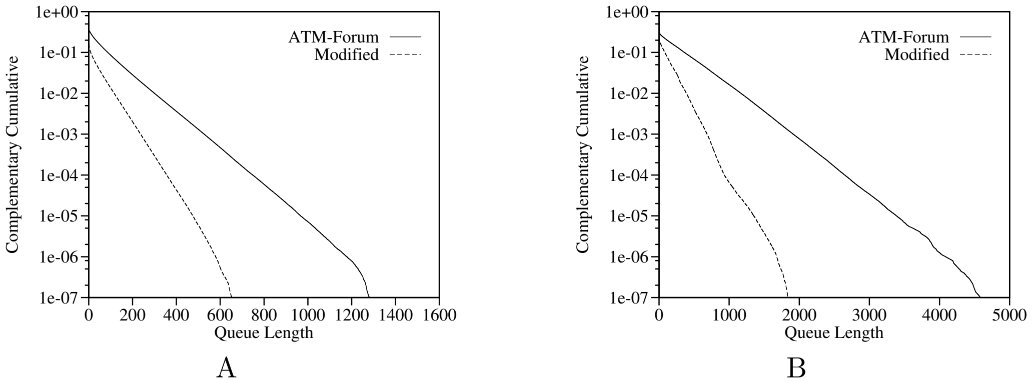


Figure 7.6: Queue length Complementary Cumulative Distribution obtained multiplexing 10 ON-OFF sources. t_a=5 ms, burst length=100 cells (A), t_a=50 ms, burst length=1000 cells (B).

mechanism, and when the sources and the switch follow our modifications. In case (A) the mean burst arrival is $ta = 5$ ms and the burst length is 100 cells. Case (B) shows the same but the burst length has been increased to 1000 cells and mean burst arrival to $ta = 50$ ms in order to maintain the same load.

Figure 7.6 shows that the queue length is considerably reduced when the sources and the switch follow our modifications. This is because the offered rate and number of active VCs are more accurately measured in our framework.

7.8 Conclusions

ER switches fairly divide the available bandwidth between the contending sources and convey the fair rate to the sources by means of the ER field of RM-cells. Sources which become silent or internally rate limited may not use this rate. To efficiently distribute the available bandwidth it is therefore important that switches keep track of the transmission rate of the sources. Using the CCR field of RM-cells to do that have the following drawbacks:

- The policing function proposed up to now does not perform a CCR conformance.
- The CCR is set to the ACR, but the ATM-Forum has not specified any rule to constrain the sources to maintain the ACR close to the transmission rate.
- A source that becomes silent stops sending RM-cells. Checking the CCR of RM-cells is therefore not enough to keep track of the source transmission rate.

In this chapter solutions to these drawbacks are proposed. A UPC which performs a CCR test is described. To encourage the sources to implement mechanisms that would maintain the ACR close to the actual transmission rate we propose charging the connections based, at least in part, on any measure of the conveyed CCR (e.g. the time average). Finally, to advertise the silent intervals, we propose that ON-OFF sources delimit activity periods by two RM-cells.

A modified ERICA switch is presented which takes advantage of our proposals. The modification consists of a simple algorithm which measures the input rate and number of active VCs by means of the conveyed CCR by forward RM-cells. This way of measuring significantly simplifies the switch mechanism. Moreover, performance results show that an ERICA switch based on our proposal gives less oscillations of the ACR of the sources and the queue length is lower. The same algorithm could be performed by other switch mechanisms.

Chapter 8

DGCRA: The Conformance Definition for ABR

8.1 Introduction

Remember from section 2.4.5 that the conformance definition is the formalism used by the network to decide whether the source transmits according to the traffic contract. The conformance definition is given by means of an algorithm which defines whether the cells passing a measuring point located at the UNI are conforming or non-conforming.

The decision of cell conformance is made by measuring the inter-cell arrival time of a connection and checking whether it deviates from the inverse of an *expected rate* a tolerance called Cell Delay Variation Tolerance (CDVT). The CDVT is also specified in the Traffic Contract and is an upper bound of the unavoidable CDV introduced by the ATM layer functions and multiplexing stages up to the measuring point.

The conformance definition may be implemented in the Usage Parameter Control (UPC) for policing. This policing function is used by a network operator to decide whether a connection is compliant or not. Furthermore, the UPC may also mark or discard non-conforming cells.

The standardization bodies have defined the Dynamic Generic Cell Rate Algorithm (DGCRA) as the conformance definition for ABR. Despite of the importance of the fact that the UPC has to guarantee the effectiveness of the network services, minor attention has been given to the DGCRA. In [78], the authors study the ability of the DGCRA to restrict ABR sources to the traffic contract by discarding excess traffic. In [47], the impact of violating over non-violating sources is investigated.

In this thesis two chapters have been devoted to the DGCRA. First, this chapter presents the state of art of the DGCRA, some improvements are proposed and a performance study is carried out based on simulation results. The next chapter studies the parameter dimensioning of the DGCRA by means of analytical models.

The rest of this chapter is organized as follows. Section 8.2 gives an overview of the ABR conformance definition. In section 8.3 improvements are proposed to the ABR conformance definition. Section 8.4 presents simulation results illustrating the performance of the improved algorithms. Finally, section 8.5 gives some concluding remarks.

8.2 The Dynamic Generic Cell Rate Algorithm (DGCRA)

The DGCRA has been defined as the conformance definition for an ABR connection by the ITU-T [35] and the ATM Forum [4]. DGCRA is an extension of the GCRA described in section 2.4.3. Basically, the increment parameter I of the GCRA (see equation (2.1)) is left variable in the DGCRA and is referred as I_k. If I_k is constant the DGCRA and the GCRA are coincident. In the following the algorithm defining the DGCRA is described.

At each cell k arrival, the DGCRA decides the cell conformance by measuring the CDV value $y_k = c_k - a_k$, where a_k is the arrival epoch and c_k is a theoretical arrival time. The cell k is non-conforming if y_k is greater than CDVT and is conforming otherwise. The theoretical arrival time is computed at each cell arrival by the following algorithm:

Let τ_1 = CDVT. After the initializations $c_0 = a_0$, $LVST_0 = a_0$, $I_0^{old} = I_0$, the set of theoretical arrival times $\{c_k\}_{k>0}$ is computed at the arrival epochs a_k as:

$$\begin{aligned}
c_k &= LVST_{k-1} + \min(I_{k-1}^{old}, I_k) \\
LVST_k &= \begin{cases} \max(c_k, a_k) & \text{if } \; y_k \leq \tau_1 \quad \text{(cell conforming)} \\ LVST_{k-1} & \text{if } \; \tau_1 < y_k \quad \text{(cell non conforming)} \end{cases} \\
I_k^{old} &= \begin{cases} I_k & \text{if } \; y_k \leq \tau_1 \quad \text{(cell conforming)} \\ I_{k-1}^{old} & \text{if } \; \tau_1 < y_k \quad \text{(cell non conforming)} \end{cases}
\end{aligned} \tag{8.1}$$

$LVST_k$ stands for the Last Virtual Scheduled Time at cell k arrival. The theoretical arrival time c_k is given by $LVST_k$ plus an increment $\min(I_k, I_k^{old})$ equal to the inverse of the expected source rate at the interface. The variable I_k follows the source rate changes expected at the interface based on the feedback conveyed by the backward RM-cell flow up to cell k arrival time a_k. I_k^{old} is introduced because the first cell received after a new increase I_k is scheduled may be received at this increase (because of the rescheduling option of the source) or at the previous increment I_k^{old}. Taking the minimum of I_k and I_k^{old} the algorithm stays on the save side.

The ATM Forum [4] gives the pseudo-code of figure 8.1 equivalent to the DGCRA given by equations (8.1). Sentences 2∼8 together with sentences 18∼19 of the algorithm of figure 8.2 perform also a conformance test to the ACR Decrease Factor (ADTF) parameter (see SES rule 5, section 4.3.1). These sentences will be explained in section 8.2.1.

Computation of sequence I_k is not an easy task because a change of rate conveyed by a backward RM-cell received at the measuring point at a given time may be applied to the forward cell flow after a delay equal to the round trip feedback from the measuring point to the source [1].

Therefore two time constants τ_2 and τ_3 are introduced to define the DGCRA which are respectively an upper bound and a lower bound of that round trip feedback delay. To be on the save side, the DGCRA schedules rate increases conveyed in the backward RM-cell flow after a delay τ_3, and rate decreases after a delay τ_2. An overview of two algorithms that have been proposed to compute I_k is given in the following section.

[1] The round trip feedback is the sum of the delay from the departure of the RM-cell from the measuring point to the receipt by the source, and the delay from the departure of the next CLP=0 cell from the source following the receipt to the arrival at the measuring point.

```
At each cell $k$ arrival epoch $a_k$ :
    if(the received cell $k$ is a CLP = 0 forward RM-cell) {
        if($a_k$ - $t_f$ > ADTF + $\tau_1$ and $ICR\_eligible$ == TRUE) {
            $PACR$ = min($PACR$, ICR) ;
            $I_k$ = 1 / $PACR$ ;
            $\boxed{der\_first = PACR ;}$
            $\boxed{der\_last = PACR ;}$
        }
        $t_f$ = $a_k$ ;
        $ICR\_eligible$ == FALSE ;
        if($CCR_k \leq PACR$) {   /* $CCR_k$ is conforming */        /* improvement 1 */
        } else /* ... */ ;      /* $CCR_k$ is non-conforming */
    }
    $c_k$ = $LVST$ + min($I_k$, $I^{old}$) ;
    if($c_k$ - $a_k \leq \tau_1$) {           /* conforming cell */
        $LVST$ = max($a_k$, $c_k$) ;
        $I^{old}$ = $I_k$ ;
    }   else /* ... */ ;          /* non-conforming cell */
```

Figure 8.1: Pseudo-code of the DGCRA. Sentences included in a box are modifications proposed in this chapter.

8.2.1 Algorithms “A” and “B” for the Determination of I_k

Two algorithms “A” and “B” have been defined for the determination of I_k ([35], [4]). Algorithm “A” provides the tightest conformance according to the delay bounds τ_2 and τ_3 described above and works as follows. Let a_k be the arrival epoch of cell k, ER_1 the last ER conveyed by a backward RM-cell previous to time $a_k - \tau_2$ and ER_2 the maximum ER value conveyed by the backward RM-cells received in the interval $[a_k - \tau_2, a_k - \tau_3]$. The increment applied to cell k should be $I_k = \min(1/ER_1, 1/ER_2)$. Because the rate must remain between PCR and MCR, the control $I_k = \max(1/\text{PCR}, \min(I_k, 1/\text{MCR}))$ has to be finally done. Note that the algorithm “A”, plus the computation of the ER_1 and ER_2 values at each arrival of a forward cell, implies storing at a given time t the rate change conveyed in the last RM-cell received previous to the time $t - \tau_2$, and all the arrival epochs and rate changes conveyed in the backward RM-cells received in the interval $[t - \tau_2, t]$. Such an algorithm would be rather complex to implement and a simpler algorithm “B” has been defined.

Figure 8.2 shows the algorithm “B” given by the ATM Forum [4] in a “C style” pseudo-code. Basically the algorithm works as follows. The variable named as $PACR$ maintains the allowed cell rate expected at the interface, i.e. at any time $I_k = 1/PACR$. $PACR$ is computed storing at most two rate changes (der_first and der_last) and their corresponding scheduled times (t_first and t_last). When the first backward RM-cell passes the interface, the first rate change der_first is scheduled at time t_first (sentences 7∼11 for an increment or 16∼17 for a decrement). Subsequent backward RM-cells that arrive before that the scheduled rate der_first takes effect (before the time reaches t_first) would modify der_first if it results in a increase of rate (sentence 7). Let denote this condition by C1. In case of a decrease the conveyed rate change would be scheduled by der_last at time t_last (sentences 13∼14) possibly overwriting a prior rate change scheduled with these variables. Let denote by C2 the case in which a scheduled rate is overwritten. When time reaches t_first, $PACR$ is updated with der_first and der_last, t_last are copied into der_first, t_first (sentences 20∼23).

Each time the condition C1 or C2 is carried out a scheduled rate change is overwritten and

```
Initialize:
        $t\_first$ = $t\_last$ = $LVST$ = 0 ; $der\_first$ = $der\_last$ = $PACR$ = ICR ;
        $I^{old}$ = 1 / ICR ; $t_f$ = INFINITY ; $ICR\_eligible$ = FALSE ;
At each backward RM-cell($ER_j$, $CI_j$, $NI_j$) departure epoch $b_j$ from the interface:
 1      $ER'$ = $ER_j$ ;
 2      if($CI_j == 1$) $ER'$ = min($ER'$, $der\_last$ (1 - RDF)) ;      /* improvement 2 */
 3      else if($NI_j == 0$) $ER'$ = min($ER'$, $der\_last$ + RIF·PCR) ;
 4      $ER'$ = min(PCR, max(MCR, $ER'$)) ;
 5      if($ER' \geq der\_first$) {              /* incr. over the first sched. rate ?  */
 6          $t\_last$ = 0 ;                      /* un-sched. future dec. rate (if any) */
 7          $der\_first$ = $der\_last$ = $ER'$ ;  /* update scheduled rate */
 8          if($t\_first < b_j$)                 /* no first rate scheduled?  */
 9              $t\_first$ = $b_j$ + $\tau_3$ ;   /* schedule new increase */
10          else if($ER' \geq PACR$)             /* reschedule if $der\_first$ was a decrease */
11              $t\_first$ = min($t\_first$, $b_j$ + $\tau_3$) ;
12      } else {                                 /* dec. over the first sched. rate */
13          $t\_last$ = $b_j$ + $\tau_2$ ;        /* schedule a decrease */
14          $der\_last$ = $ER'$ ;
15          if($t\_first < b_j$) {               /* no first rate scheduled?  */
16              $t\_first$ = $t\_last$ ;          /* copy the last event into the first */
17              $der\_first$ = $der\_last$ ;
            }
        }
18      if($b_j$ + $\tau_2$ < $t_f$ + ADTF - $\tau_1$) $ICR\_eligible$ = TRUE ;
19      else $ICR\_eligible$ = FALSE ;
When time reaches $t\_first$:
20      $PACR$ = $der\_first$ ;                  /* update $I_k$ with the first sched. rate */
21      $I_k$ = 1 / $der\_first$ ;
22      $t\_first$ = $t\_last$ ;                  /* copy the last event into the first */
23      $der\_first$ = $der\_last$ ;
```

Figure 8.2: Algorithm "B" for the calculation of I_k. Sentences included in a box are an improvement proposed in this chapter.

will not take place resulting in a decrease of the tightness of the rate conformance. Clearly the higher the value of τ_2 and τ_3 and the rate at which backward RM-cells are received, the higher will be the number of RM-cells which will cause the condition C1 or C2 to happen and therefore the less accurate will be the algorithm.

The sentences 2∼8 of figure 8.1 together with sentences 18∼19 of figure 8.2 perform a conformance test to the ADTF parameter. The auxiliary variable t_f keeps the arrival time of the last forward RM-cell received (sentence 7, figure 8.1). The condition 2, figure 8.1, decides that the source should have performed the ADTF adjustment when the time interval between two consecutive forward RM-cell arrivals is higher than the ADTF value plus the CDVT, and the auxiliary variable *ICR_eligible* is TRUE. The sentence 18, figure 8.2, sets *ICR_eligible* to TRUE just in case the algorithm has the certainty that the last backward RM-cell that passed the interface arrived to the source before the possible ADTF adjustment. Note that such backward RM-cell arrival could modify the ACR of the source set to the ICR in the ADTF adjustment.

The sentence 5 of figure 8.1 has been introduced motivated by the following reasoning (motivation of sentence 6 is given in section 8.3.3). The condition of sentence 2, figure 8.1, and sentence 18, figure 8.2, yields that $a_k > b_j + \tau_2 + 2\tau_1$ which implies that $a_k > t\_first$ (note that t_first can at most hold the value $b_j + \tau_2$), and thus that the last scheduled rate by der_first has already taken place. When the ADTF adjustment is performed by the conformance algorithm, *PACR* is set to a value probably different than der_first. Consequently, when the first backward RM-cell is received by the algorithm after the ADTF adjustment, it may be interpreted as a decrease of rate (the condition of sentence 5, figure 8.2 would not be satisfied) and accordingly der_first may be scheduled at time $b_j + \tau_2$. However, if the ER value conveyed by such backward RM-cell is greater than *PACR*, this should be interpreted as an increase of rate and scheduled at the earlier time $b_j + \tau_3$.

8.3 Improvements of the DGCRA

8.3.1 CCR-Conformance

In an ABR connection each time a source transmits an RM-cell, it sets the CCR field to the last computed ACR. CCR may be used by the switches as an approximate value of the transmission rate of the source, and several switch mechanisms use it in the computation of the ER feedback (e.g. ERICA [41] and EPRCA [68]).

At any time the CCR conveyed by a forward RM-cell received at the measurement point should not be higher than the expected rate computed at the UPC (referred as *PACR* in the algorithm of figure 8.1). Transmission at a CCR higher than the ACR would imply exceeding the permitted rate. Furthermore, if the CCR is erroneously set to a value different than the ACR, it could result in a misbehavior of the feedback control of switches that make use of the CCR.

We propose the sentences labeled *improvement 1* in figure 8.1 to check that the CCR is not higher than the expected rate *PACR*. These sentences would prevent a source from setting the CCR to a value higher than the ACR. Note however that performing a CCR-conformance test would also imply checking that the CCR is not set to a lower value than the ACR. If this happens, switches that use the CCR to estimate the transmission rate of the source would compute a rate lower than the actual transmission rate of the source. In the following section a DGCRA based on the conveyed CCR which would overcome this problem is proposed.

8.3.2 A DGCRA Based on the Conveyed CCR

As stated in section 8.2.1, the tightness of the rate conformance of the DGCRA may be reduced when used together with the algorithm "B". Moreover, the absence of an accurate CCR-conformance may lead to a misbehavior of the feedback control of switches that make use of the CCR if any source does not properly sets the CCR to the ACR. This problem was already introduced in chapter 7, where a UPC based on the CCR was introduced. Now, the same idea is revisited and extended to the DGCRA. The modifications that this proposal would require to introduce into the ABR source behavior given by the ATM Forum were given in section 7.3. For sake of completeness, these modifications have been also included in the following description.

Description of the Proposal

If the DGCRA would police a rate equal to the CCR conveyed by the forward RM-cells (e.g. adding $I_k = 1/\text{CCR}$ after the sentence 9, figure 8.1 and removing sentence 21, figure 8.2), the previous drawbacks could be overcome. However, with the current source behavior given by the ATM Forum a source may transmit cells following a forward RM-cell at a rate up to the CCR until a backward RM-cell conveying a different rate is received. Consequently, with such source behavior, at the interface it is not possible to update I_k with all the CCR values because at any time cells may be received at a higher rate.

One possibility which would permit updating I_k with all the CCR conveyed values would arise if the source was constrained to delay the rate increases just after a forward RM-cell transmission. This RM-cell would carry the new rate increase into the CCR field (i.e. cells arriving at a higher rate will be preceded by an RM-cell conveying the new rate). Including this condition in the source behavior the ABR conformance could be performed at two levels. At the first level the algorithm would check that the source correctly changes the ACR by means of a CCR-conformance. This would be performed by computing the *PACR* with the "B" algorithm of figure 8.2 and checking the conformance condition 9~10 in figure 8.1. At a second level cells would be checked by the DGCRA to be conforming to the conveyed CCR values, by updating I_k with such values.

However, remember that the RM-cells are transmitted after each (Nrm-1)th data-cell. Therefore, if the source ACR is low the time interval between RM-cells could be too large, slowing down the ramp-up of the sources and consequently the access to the unused network bandwidth. This drawback could be overcome if, when receiving a rate increase, the source was allowed to advance an RM-cell transmission before its Nrm-th turn. However, this would increase the RM-cell rate and thus the overhead introduced by the RM-cell transmission. Therefore, an "RM-cell advance algorithm" should be performed in order to decide which rate increases are worth to advance an RM-cell transmission. Note also that if the RM-cell advance is not performed using the binary feedback, a burst of backward RM-cells with CI=0 could be received while delaying the rate increases in the forward flow after an RM-cell transmission. This could result in a excessive rate increase.

A simple RM-cell advance algorithm would be the following. Let PACR be the rate increase conveyed by a backward RM-cell. The algorithm would consist of advancing a forward RM-cell transmission if PACR $\geq \alpha \cdot$ ACR, $\alpha > 1$. Clearly, the higher the value of α, the higher the rate increase required to perform the RM-cell advance, and thus the less times such advance will be performed. Note that setting $\alpha = 1$ would imply performing the RM-cell advance at any rate increase.

An Implementation Example

Figure 8.3 shows a simplified flow chart of the source behavior given by the ATM Forum (confront with figure 4.4), including the modifications previously described. In figure 8.3 the source rate is ACR, but the auxiliary variable PACR is introduced to carry out all rate changes. Note from the SES RECEPTION that if a backward RM-cell conveys a rate decrease the ACR is set to the PACR, but if it results in a rate increase the ACR is not changed. Note also from the SES & DES TRANSMISSION that previously to an RM-cell transmission, the source sets ACR = PACR. Therefore, rate increases will be delayed just after a forward RM-cell transmission. In the SES RECEPTION it can be seen that the RM-cell advance is performed when PACR $> \alpha \cdot$ ACR. Note that when this condition is accomplished Counter is set to Nrm which implies that the next cell transmission will be an RM-cell. After an RM-cell advance the rescheduling option can also be performed.

Finally, note that the former proposal would require that the order of the forward RM-cells relative to the other cells up to the interface is not changed. By now, the ATM Forum standard [4] establishes that switches may transmit RM-cells out of sequence with respect to data cells in

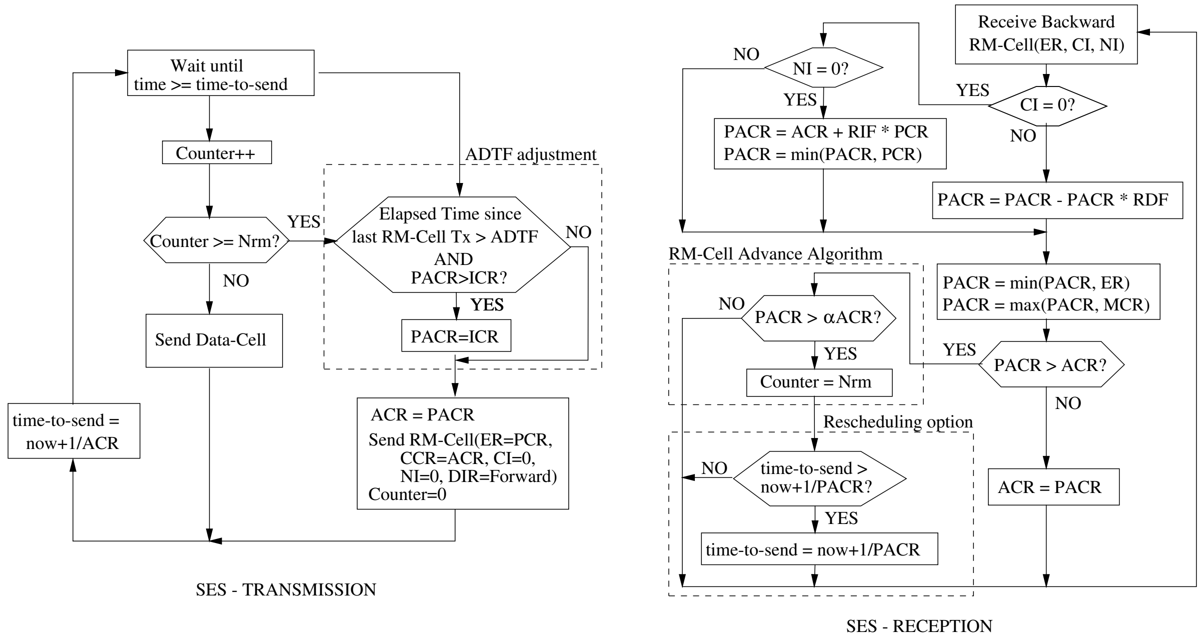


Figure 8.3: Modified source behavior which allows a DGCRA based on the conveyed CCR.

order to speed-up the feedback control. In any case, performing this change of sequence prior to the interface would imply rescheduling the cell transmission in order to maintain the inter-cell time according to the rate and CDVT values, which would increase the complexity of the switch design. Another advantage of maintaining the order of the cell flow up to the measuring point is the simplification of the RM-cell compliance measure (cfr. [4, appendix III.4]).

8.3.3 CI and NI Conformance

Algorithm “B” as defined by the ITU-T [35] and the ATM Forum [4] just takes into account the feedback conveyed by the ER of the backward RM-cells (thus defines a so called ER-conformance definition). However to specify a conformance definition to the SES behavior given by the ATM Forum [4] (see section 4.3), feedback conveyed by the CI and NI bits of backward RM-cells should be also considered. For example, an ER-conformance definition will be inappropriate for a binary switch which conveys the congestion information by means of the CI bit.

In this section a modification of the algorithm “B” is proposed in order to extend the conformance to CI and NI. The modification proposed here would also improve the ER-conformance in the following way. If the rate conformance is based exclusively on the ER feedback, it doesn’t take into account the ACR increases limited by the RIF parameter. An EPRCA switch for example [68], just modifies the ER field of backward RM-cells in case of congestion. This means that during the non-congested periods the algorithm would control a source rate equal to the PCR (value conveyed by the ER field), while the source would be limited by the rate increases given by the RIF parameter. Another example is the ramp-up period of a source controlled by an ERICA switch. This switch modifies always the ER field of backward RM-cells but in the ramp-up it is likely to happen that increases limited by the RIF parameter are more restrictive than the conveyed ER.

To solve these drawbacks we propose the sentences labeled as *improvement 2* in figure 8.2. The

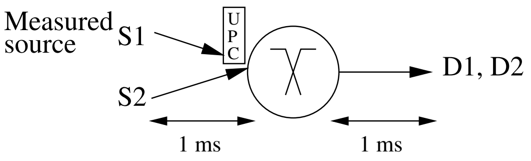


Figure 8.4: Model used in the simulations.

Sources						Switch		
PCR (Cells/ms)	MCR (Cells/ms)	ICR (Cells/ms)	RIF	RDF	Nrm	EFCI	ERICA	
						Qth	TCR	N
365	2	10	1/64	1/16	32	100	0.85 LCR	30

Table 8.1: Sources and switch parameters.

following considerations motivate these sentences. The *der_last* variable of the algorithm keeps track of the rate changes conveyed by all the backward RM-cells that pass the interface. Our proposal consists of using *der_last* to compute the new rate at which the source may change the ACR value depending on the CI and NI bits of the backward RM-cell. Doing this the *der_last* variable will carry out an upper bound of the rate changes accomplished by the ACR value of the source. In order to follow all the source rate changes with the *der_last* variable, the initialization of sentence 6, figure 8.1 needs to be perform, otherwise *der_last* would not follow the rate decreases down to ICR performed in the ADTF adjustment.

Applicability of the Improved Algorithm for the UPC of Binary Switches

If the rate changes are conveyed exclusively by the CI and NI bits (as in the binary switches), the rate at a given instant may depend on all the previously computed rates. This makes the policing of sources controlled by binary switches a rather difficult task. Assume for example a source multiplexed by binary switches and a UPC performing the CI and NI conformance previously described. If the source computes a lower rate than the UPC when receiving a backward RM-cell, the following rate changes computed at the UPC would overestimate the ACR value which should be policed. Note that using the ER-feedback, each conveyed ER may carry the absolute value at which the source sets the ACR and thus "synchronizes" the source and the UPC and avoids this problem.

Several reasons may lead a source to compute a lower rate than the UPC using the algorithm described in section 8.3.3. First of all, a congested switch located between the source and the UPC would introduce rate decreases which could not be taken into account by the algorithm. Secondly, the source may carry out rate decreases that are not performed by the former algorithm, as the CRM and the take-or-lose-it adjustments (see section 4.3). Another reason is that the ATM Forum establishes that the source must adjust its rate to a value lower or equal to the conveyed rate by the backward RM-cells.

We conclude that in order to use the improved algorithm as a UPC for binary switches, the UPC should check that the source and the algorithm perform the same rate changes and "re-synchronize" both in case the expected rate is overestimated.

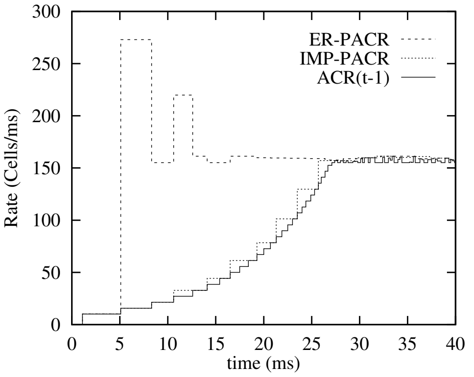


Figure 8.5: ERICA switch. In the ramp up, rate is limited by the RIF.

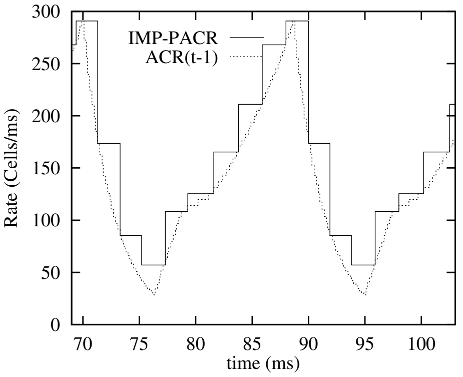


Figure 8.6: Binary switch. The improved algorithm "B" performs a CI-conformance.

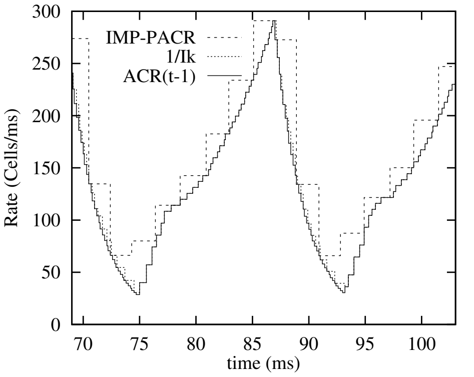


Figure 8.7: Binary switch. The DGCRA based on the conveyed CCR performs a tight rate conformance ($1/I_k$).

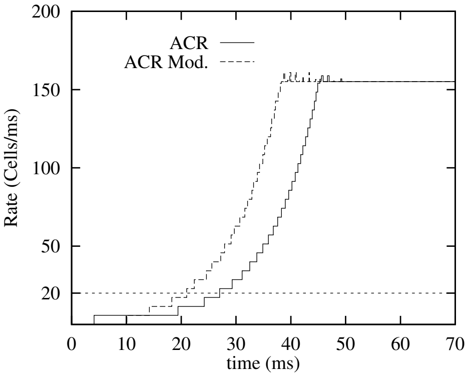


Figure 8.8: ERICA switch. The RM-cell advance algorithm speeds-up the rate ramp-up.

8.4 Performance Results

Figure 8.4 shows the network configuration used in the simulations. Two ABR greedy sources[2] feed an EFCI/ERICA switch with parameters given in table 8.1. The link capacity have been set to LCR = 365 Cells/ms (~ 155 Mbps). The ABR sources behave as described in section 4.3, and do not implement the rescheduling option. One of the sources cell flow passes a UPC implementing the DGCRA with the algorithm "B" described in section 8.2. In our model no CDV is introduced in the cell flow up to the UPC, therefore the feedback delay is equal to the round trip propagation delay (2 ms), and thus the values $\tau_2 = \tau_3 = 2$ ms have been set in the conformance algorithm of the UPC. We are just interested in the ability of the algorithm to compute an accurate value of the expected rate (PACR), therefore the parameter τ_1 has no influence in our results. The figures that will be presented below compare the PACR and the rate at which the forward cell flow arrives at the interface, which will be the transmission rate (ACR) delayed the propagation delay (1 ms), thus ACR(t-1) has been plotted. PACR would be coincident with ACR(t-1) in a UPC performing the tightest rate conformance.

[2] An ABR source is called "greedy" when has always cells to send and transmits at the maximum permitted rate i.e. the ACR.

Figure 8.5 shows the ACR(t-1) of the measured source, the expected rate (PACR) computed by a UPC using the conformance definition with the algorithm “A” given by the ATM Forum [4] (indicated as ER-PACR) and the same value (indicated as IMP-PACR) computed by a UPC with the *improvement 2* proposed in this chapter (see section 8.3). The switch is of the ERICA type. The figure shows how, in the ramp-up, the ACR of the source is limited by the RIF, and the improved algorithm yields a better conformance test.

Figure 8.6 shows the ACR(t-1) of the source and the PACR of the improved algorithm with the binary EFCI switch. Both curves are not coincident because the algorithm “B” keeps at most two scheduled rates (see section 8.2.1). The result of this simplification is that the PACR is a step envelope of the ACR(t-1) with a step duration nearly equal to the feedback delay. Consequently, the higher the feedback delay is, the less tight the rate conformance is.

Finally, figures 8.7 and 8.8 show the performance that would be obtained with the DGCRA based on the conveyed CCR proposal of section 8.3.2. In order to set the parameter α of the advance algorithm the following reasoning can be done. During the non congested periods, in the ramp-up the sources are allowed to increase their rate in steps given by the RIF, resulting $\text{PACR} \geq \text{ACR} + \text{RIF} \cdot \text{PCR}$. During such periods we have set the condition of performing the RM-cell advance until a certain rate L is reached. Note that when the rate L is reached, the maximum delay between the RM-cell transmission will be Nrm/L. $L = 20$ cells/ms has chosen which, for $\text{Nrm} = 32$ cells, gives a maximum delay of 1.6 ms. The RM-cell advance is performed when $\text{PACR} \geq \alpha \cdot \text{ACR}$. Therefore, from the previous relations we derive that α should be set to $\alpha = 1 + \text{RIF} \cdot \text{PCR}/L$. Setting the source parameters of table 8.1 and $L = 20$ cells/ms we get $\alpha = 1.28$.

In figure 8.7 the same simulation as in figure 8.6 has been performed, but with the source behavior and the UPC algorithm proposed in section 8.3.2. Note that now the rate conformance performed at cell level is carried out using I_k, which is updated with the conveyed CCR, resulting a very tight conformance test. Remind that the PACR is used to check whether the source correctly updates the CCR.

In order to see the performance of the proposed RM-cell advance algorithm, the sources have been modified in figure 8.8 to start with a low rate, setting $\text{MCR} = \text{ICR} = 0.1$ cell/ms. Note that the interval between the RM-cell transmission at this rate without the RM-cell advance would be 320 ms (with $\text{Nrm} = 32$). As explained above, α has been set in order to have an RM-cell advance until the rate $L = 20$ cells/ms is reached. Figure 8.8 shows the rate of the measured source obtained in two simulations: one in which the sources follow the ATM-Forum behavior (ACR), and another in which they follow the source behavior proposed in section 8.3.2 (ACR Mod.). Figure 8.8 shows that delaying the rate increases until an RM-cell transmission does not slow down the ramp-up because of the RM-cell advance. Moreover, the RM-cell advance produces a speed-up of the ramp-up over the sources which follow the ATM-Forum behavior. The overhead of transmitting more RM-cells is very low. The source following the modified behavior just advances 4 RM-cells (transmitted before the rate $L = 20$ is reached).

8.5 Conclusions

This chapter describes the DGCRA and the algorithm “B” given by the ATM Forum [4] as an example conformance definition of the ABR Service. Algorithm “B” has the drawback of decreasing the accuracy of the rate conformance with increasing feedback delay.

We have proposed a first improvement of the DGCRA which consist of checking the CCR field to be lower than the expected rate as part of the conformance test, i.e. to perform a CCR conformance. Non-conforming CCR would imply that the allowed cell rate of the source exceeds the permitted rate and could result in a misbehavior of the feedback control of switch mechanisms that use the CCR in the computation of the ER of backward RM-cells.

To achieve a tight rate conformance we have extended the former improvement proposing a conformance definition based on the conveyed CCR. This would imply modifying the source behavior to delay the ACR increases just after a forward RM-cell transmission which would convey the new rate into the CCR field. To avoid an excessive delay of rate increases, we have proposed a simple RM-cell advance algorithm.

A second improvement is proposed that would extend the algorithm for CI and NI conformance. Furthermore, the applicability of the improved algorithm for the UPC of binary switches is discussed.

Performance results carried out by simulation of two switch mechanisms (EFCI and ERICA), show the effect of feedback delay on the tightness of the rate conformance of the algorithm “B”, and the effectiveness of the proposed improvements.

Chapter 9

DGCRA Parameter Dimensioning

9.1 Introduction

In the previous chapter the conformance definition for ABR (the DGCRA) was described and the possible problems of the algorithm were analyzed. In this chapter the DGCRA is studied again, but from the point of view of the parameter dimensioning.

Remember from section 8.2 that three parameters used by the DGCRA are negotiated at the connection setup: a Cell Delay Variation Tolerance (CDVT), and upper and lower bounds of the round trip delay between the UPC and the source (referred to respectively as τ_2 and τ_3). In this chapter the dimensioning of these parameters is analyzed.

In this chapter a "well behaving source" is assumed and the effect of tuning the different DGCRA parameters is investigated on the rejection probability. This is done by modeling the DGCRA as a queue and solving the model by means of two analytical approaches: one based on a renewal assumption and another one using a matrix geometric technique.

The effect that the CDV in the ABR cell stream may have in the network stability is also investigated. This is done by means of an ABR "worst case traffic" which transmits bursts of back to back cells up to the negotiated CDVT.

The rest of the chapter is organized as follows. In section 9.2 an equivalent queuing model for the DGCRA is described. In section 9.3 this model is solved by means of a renewal approximation approach. Results obtained by simulation are shown to obtain CDVT dimensioning guidelines for ABR, and the analytical method is applied and confronted with the simulation results. In section 9.4 the queuing model of the DGCRA is solved using a matrix geometric approach. Numerical results are also shown and validated by simulation. In section 9.5 the effect of the CDVT is investigated by means of an ABR "worst case traffic". Finally, in section 9.6 some concluding remarks are formulated.

9.2 Queuing Model

The GCRA (see section 2.4.3) can be modeled as a single server queue with a workload (unfinished work) limited to the CDVT (see e.g. [66]). This queuing model can be extended to the DGCRA conformance definition given by equations (8.1). In the model we make the following

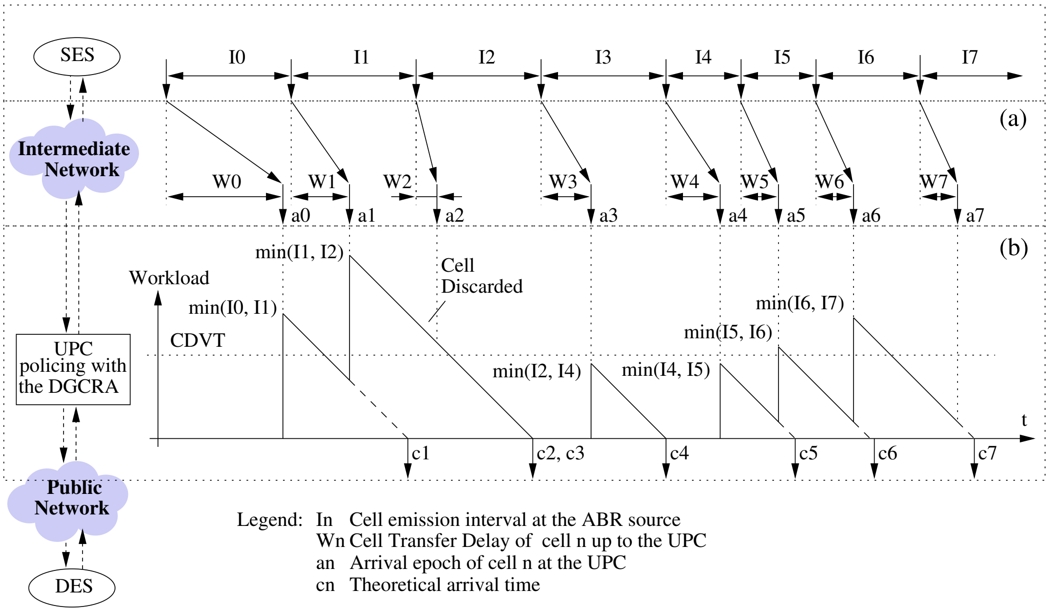


Figure 9.1: CDV introduced by an intermediate network (a) and the workload at the equivalent queuing model of the DGCRA (b).

assumptions: (i) the same rate changes followed by the source are policed by the DGCRA (ii) The DGCRA is able to properly schedule the increments I_n corresponding to these rate changes (iii) No rescheduling option is used by the source. With these assumptions, when cell n arrives at the DGCRA the increment I_n is equal to the source emission interval between cell n and cell $n+1$. This will be our definition of I_n in the rest of the chapter.

The behavior of the queue which we use to model the DGCRA is shown in the time diagram of figure 9.1. In this queuing model the cells correspond to customers. The service time is the increment applied by the DGCRA and the theoretical arrival time of cell n is the departure time of the previous accepted cell (c_n in the figure). The CDV value y_n of equations (8.1) is given by the workload of the queue when $y_n > 0$. Therefore, when a cell arrival finds a workload higher than the CDVT, a non-conforming condition is given.

Note that the increment added at the cell n arrival epoch a_n in (8.1), is the service time to be added at the cell arrival epoch a_{n-1} in the equivalent queuing model. In other words, the workload increment added at the arrival epoch a_n in the queuing model is the increment that would be added by the DGCRA at the cell arrival epoch a_{n+1}, and is therefore given by $\min(I_n^{old}, I_{n+1})$ if cell $n+1$ is accepted, and 0 otherwise. For sake of simplicity, in figure 9.1 we have approximated this by adding $\min(I_n^{old}, I_{n+1})$ to the workload if the cell n is accepted and 0 if cell n is not accepted.

9.3 Renewal Approximation Approach

M. Ritter and P. Tran-Gia analyze the GI/GI/1 queue with bounded delay in [65]. They do a discrete time analysis where the time slot is equal to the transmission time of one cell. The authors apply the method to compute the rejection probabilities of a cell stream policed by the

GCRA. In this case the service time is deterministic. In the following we extend the method to compute the rejection probabilities for the DGRCA. Now the service time is not deterministic since the emission interval of the cells at the ABR source is not constant.

The DGCRA is modeled by means of the queue described in the previous section. For sake of simplicity we assume that the increment $\min(I_n^{old}, I_{n+1})$ of this queuing model is the emission interval of the last accepted cell. This is a worst case assumption since in practice the UPC may underestimate the inter-cell of the emitting source for several reasons. First because the UPC takes the $\min(I_n^{old}, I_{n+1})$. Secondly because the algorithm used by the UPC to determine the increments may not be able to follow all the rate changes conveyed by the backward RM-cells (see chapter 8 for details). Finally, because the source may transmit at a lower rate if it is internally rate limited or if there is a bottleneck switch between the source and the UPC.

Let U_n be the unfinished work of the queue at the n cell arrival. We denote the CDVT by τ_1 [1]. For convenience, the following random variables are defined:

$$U_{n,0} = U_n | U_n \leq \tau_1, \qquad U_{n,1} = U_n | U_n > \tau_1 \tag{9.1}$$

$$U_{n+1,0} = U_{n+1} | U_n \leq \tau_1, \quad U_{n+1,1} = U_{n+1} | U_n > \tau_1 \tag{9.2}$$

Given the probabilities $u(k)$ of U_n ($u(k) = \text{Prob}\{U_n = k\}$), the probabilities $u_{n,0}(k)$ and $u_{n,1}(k)$ of $U_{n,0}$ and $U_{n,1}$ are given respectively by:

$$u_{n,0}(k) = \frac{\sigma^{\tau_1}[u(k)]}{\text{Prob}\{U_n \leq \tau_1\}} \tag{9.3}$$

$$u_{n,1}(k) = \frac{\sigma_{\tau_1+1}[u(k)]}{\text{Prob}\{U_n > \tau_1\}} \tag{9.4}$$

where:

$$\sigma^{\tau_1}[x(k)] = \begin{cases} x(k), & k \leq \tau_1 \\ 0, & k > \tau_1 \end{cases}$$

$$\sigma_{\tau_1+1}[x(k)] = \begin{cases} 0, & k < \tau_1 + 1 \\ x(k), & k \geq \tau_1 + 1 \end{cases}$$

According to the time diagram of figure 9.1, the unfinished work U_n is related to the inter-cell I_n and the inter-arrival A_n at the UPC by:

$$\begin{array}{ll} U_n \leq \tau_1, & U_{n+1,0} = \max\{0, U_{n,0} + I_n - A_n\} \\ U_n > \tau_1, & U_{n+1,1} = \max\{0, U_{n,1} - A_n\} \end{array} \tag{9.5}$$

We define $D_n = W_{n+1} - W_n = -(I_n - A_n)$. Note that D_n is the cell transfer delay variation of two consecutive cell arrivals at the UPC (cfr. figure 9.1). Assuming that the random variables

[1] This notation is used in the ATM Forum standard and we will indistinctly use CDVT and τ_1 in the rest of the chapter.

A_n are independent and identically distributed with probabilities $a(k)$, and independent of U_n, and assuming the same for D_n with probabilities $d(k)$, from equations (9.5) we have:

$$u_{n+1,0}(k) = \pi_0\left[u_{n,0}(\cdot) * d(-k)\right] \tag{9.6}$$
$$u_{n+1,1}(k) = \pi_0\left[u_{n,1}(\cdot) * a(-k)\right] \tag{9.7}$$

where $\pi_0\left[x(k)\right] = \begin{cases} 0, & k < 0 \\ \sum_{i=-\infty}^{0} x(i), & k = 0 \\ x(k), & k > 0 \end{cases}$

and $u(\cdot) * d(-k) = \sum_{l=-\infty}^{\infty} u(l)\, d(l-k)$.

Note that $u_{n+1}(k) = \text{Prob}\left\{U_n \leq \tau_1\right\} u_{n+1,0}(k) + \text{Prob}\left\{U_n > \tau_1\right\} u_{n+1,0}(k)$, therefore, from the previous equations we obtain the following recursive relation:

$$u_{n+1}(k) = \pi_0\left[\sigma_1^{\tau}\left[u_n(\cdot)\right] * d(-k) + \sigma_{\tau_1+1}\left[u_n(\cdot)\right] * a(-k)\right] \tag{9.8}$$

The first cell arrival at the UPC encounters $U_0 = 0$. Therefore, given the inter-arrival $a(k)$ and the cell transfer delay variation $d(k)$ histograms at the UPC, we can obtain the distribution of the unfinished work $u_n(k)$ applying the following iterative algorithm:

- Initialize $u_0(k) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$
- Iterate equation (9.8) until the equilibrium probabilities $u(k) = \lim\limits_{n\to\infty} u_n(k)$ are reached

Finally, the rejection probability is given by:

$$\text{Rejection probability} = \sum_{i=\tau_1+1}^{\infty} u(i) \tag{9.9}$$

The histograms $a(k)$ and $d(k)$ can be obtained, for example, by simulation or measurements.

9.3.1 Numerical Results

In this section we investigate the CDVT dimensioning of ABR by simulation, we show the usage of the iterative method described in section 9.3 and we validate its applicability.

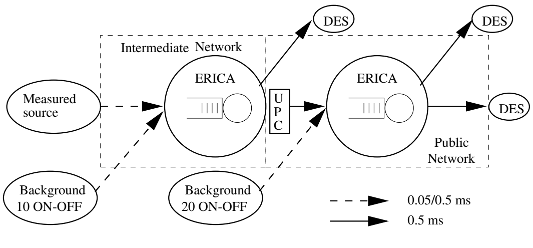


Figure 9.2: Network Topology.

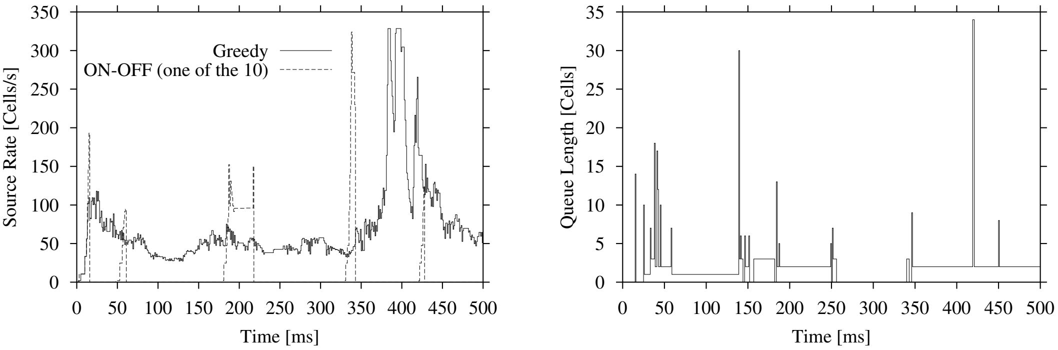


Figure 9.3: Rate of the greedy source and one of the ON_OFF sources and queue length in the LAN framework.

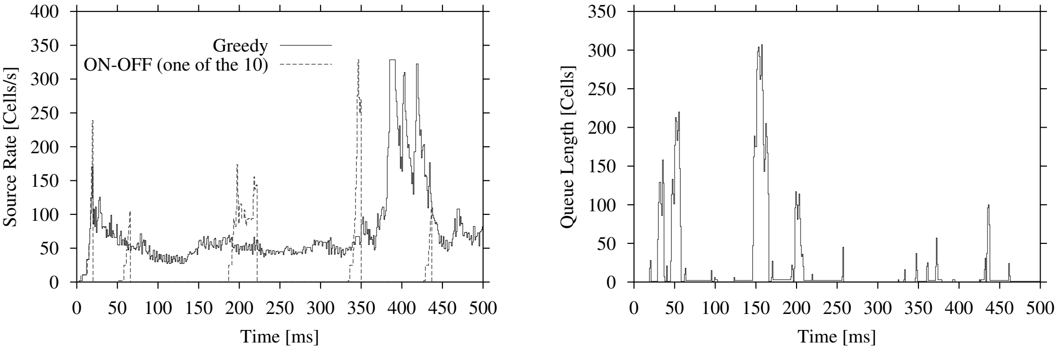


Figure 9.4: Rate of the greedy source and one of the ON_OFF sources and queue length in the MAN framework.

We use network topology of figure 9.2. The system consists of an ABR greedy source that is multiplexed at each switch with a background of ABR ON-OFF sources which behave as follows. During the ON period a burst of cells is transmitted with greedy behavior. The number of cell of the bursts is geometrically distributed with a mean of 2000 cells. After the last cell is transmitted, the source remains silent for an exponentially distributed time with a mean of 100 ms and then a new burst is transmitted. The source sets the ACR to the ICR at the beginning of each burst transmission. In the simulation we have used the emission intervals of the ABR source as the increments in the DGCRA, like it was assumed in the queuing model described in section 9.2.

For the ABR sources the parameters RIF = 1/16 and ICR = 2 cells/ms were fixed. The switch implements the ERICA algorithm (see section 5.2.3). The parameters of the switch are: Utilization Factor = 0.9, Counting Interval in cells = 100, and the Max-Min Fairness option was used.

Two link delays from the sources to the switches were considered: 0.05 ms and 0.5 ms. We will refer to these delays as the LAN and the MAN framework respectively. The other link delays were fixed to 0.5 ms. The link rate was set to 365 cells/ms (≈ 155 Mbps).

Figures 9.3 and 9.4 show the evolution of the source rates (the greedy source and one of the

ON-OFF sources) and the queue length of the intermediate network switch in the LAN and the MAN framework respectively. The figures just show the first 500 ms of the simulation. Longer links delays imply a slower reaction time, this motivates the higher values of the queue length observed in the MAN framework.

Figure 9.5 shows the variable component of the Cell Transfer Delay (CTD) probabilities up to the UPC in the LAN and MAN framework. In the LAN framework the CTD reaches lower values than in the MAN framework because the queue length is lower. Figure 9.6 shows the Cell Delay Variation (CDV) probabilities defined as the histogram of $y_n = c_n - a_n$ when CDVT $\to \infty$, and thus none of the cells is discarded. From figure 9.6 we can see that a CDVT $\approx$ 95 time-slots and CDVT $\approx$ 475 time-slots would be required in the LAN and the MAN frameworks respectively to avoid a non-conformance cell condition.

Figures 9.7 and 9.8 show the probabilities $a(k)$ and $d(k)$ to be used by the iterative method described in section 9.2. Remember that $a(k)$ and $d(k)$ are respectively the cell inter-arrivals and the Cell Transfer Delay Variation (CTDV) probabilities at the UPC. Note that although the CTD are on the average about one order of magnitude higher in the MAN than in the LAN framework (cfr. figure 9.5), the inter-arrivals and the CTDV remain with little changes in both frameworks (cfr. figures 9.7 and 9.8).

The probabilities $a(k)$ and $d(k)$ shown in figures 9.7 and 9.8 have been used to compute the rejection probabilities applying the algorithm described in section 9.2. Figure 9.9 shows the evolution of the unfinished work for different number of iterations. The curves have been obtained iterating equation (9.8) using the probabilities $a(k)$ and $d(k)$ obtained in the MAN framework for a CDVT of 50 and 250 time slots.

Figure 9.9 shows the convergence of the algorithm described in section 9.2. Note that the unfinished work probabilities converge to a nearly uniform distribution between 0 and the CDVT. This is reasonable since the load of the equivalent queue is 1 until cells are discarded (when the arriving cell finds an unfinished work greater than the CDVT). Note also that the higher the CDVT, the higher the number of iterations required by the unfinished work probabilities to converge. Furthermore, the higher the CDVT, the higher the number of operations required at each iteration, since the sequence is longer. Consequently, the speed of the algorithm decreases very fast with increasing CDVT.

Finally, figure 9.10 shows the rejection probabilities obtained with the iterative method and with simulation for different CDVT values. Since the probabilities $a(k)$ and $d(k)$ have little differences in the LAN and the MAN framework, the rejection probabilities obtained with the iterative algorithm are nearly the same. On the other hand, using simulation we observe that in fact the rejection probabilities are much higher in the MAN than in the LAN framework. This is motivated by the higher queue lengths that are built up in the MAN framework. This is reflected in figures 9.5 and 9.6 where the CTD and consequently the CDV are much higher in the MAN than in the LAN framework. With a higher CDV, a higher CDVT is required to maintain a low rejection probability.

Moreover, figure 9.10 shows that the rejection probabilities decrease much more slowly with the iterative algorithm than by simulation. Note from this figure that by simulation the rejection probability decreases nearly exponentially with increasing CDVT. However, in the iterative algorithm we have observed that the rejection probability decreases proportionally to 1/CDVT. This is because the unfinished work probabilities in the iterative algorithm have a nearly uniform distribution between 0 and the CDVT, as shown in figure 9.9.

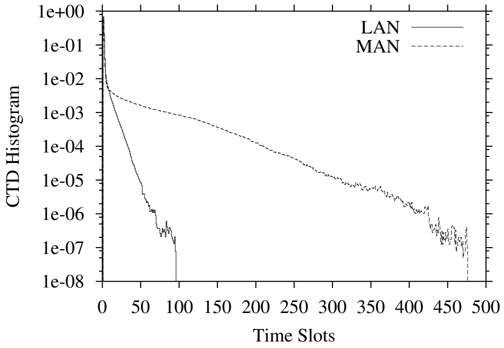


Figure 9.5: Histogram of the variable part of the cell transfer delay.

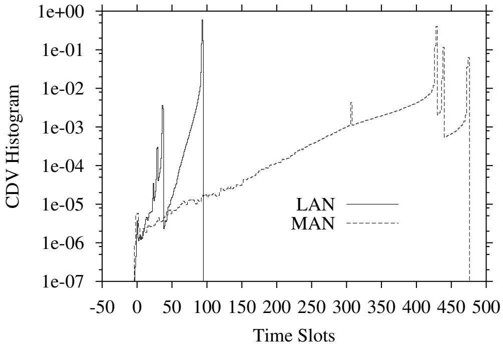


Figure 9.6: Histogram of the cell delay variation ($y(k)$).

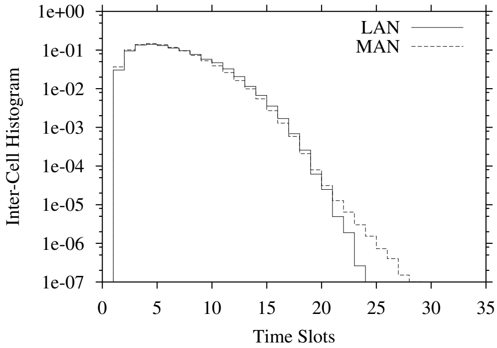


Figure 9.7: Histogram of the cell inter-arrivals at the UPC ($a(k)$)

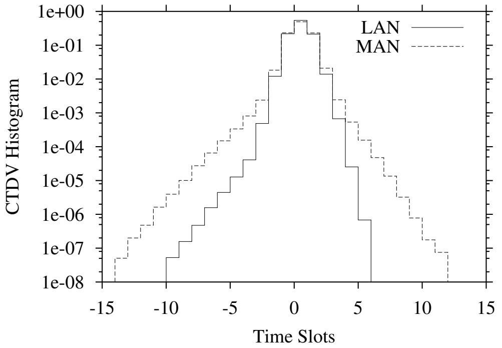
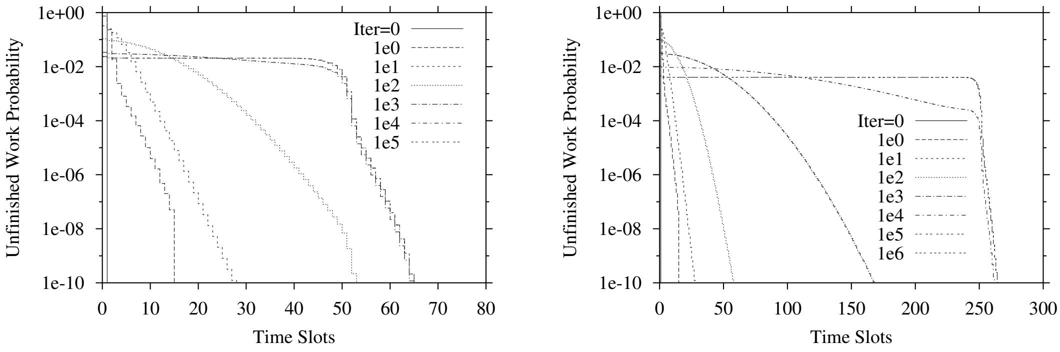


Figure 9.8: Histogram of the cell transfer delay variation ($d(k)$).



CDVT = 50 time slots

CDVT = 250 time slots

Figure 9.9: Evolution of the unfinished work increasing the number of iterations.

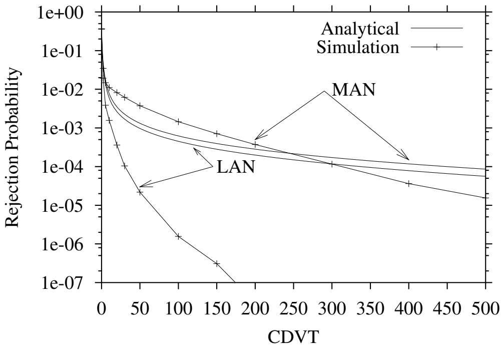


Figure 9.10: Rejection probability obtained analytically and by simulation.

9.4 Matrix Geometric Approach

In this section a matrix geometric approach is used to evaluate the DGCRA. This analytical method only allows to model simple network topologies, otherwise the size of the matrices involved in the method grow too much and cannot be solved. Therefore, the numerical results obtained in this section have been obtained with a simpler network topology than in section 9.3. However, this method gives more accurate results and has been used to investigate the influence of the delay bounds τ_2 and τ_3. This has been done considering two scenarios which allow to capture the case when these bounds are properly or not properly set.

The rest of the section is organized in the following subsections. First, subsection 9.4.1 presents a detailed description of the analytical model used in the evaluation. The two scenarios used to investigate the delay bounds τ_2 and τ_3 are described in subsection 9.4.2. The model is solved in section 9.4.3. Finally, subsection 9.4.4 gives numerical results obtained from the analytical model.

9.4.1 Model Used in the Evaluation

In order to evaluate the DGCRA we do a discrete time analysis of the system shown in figure 9.11. In this system a single ABR source is multiplexed with a VBR source. The VBR source has full priority over the ABR source. This multiplexing stage models the jitter introduced in the intermediate network shown in figure 9.1.

The assumptions made in the equivalent queue of the DGCRA described in section 9.2 apply to our model. Note that these assumptions imply that the ABR source generates as much traffic as possible without violating the allowed cell rate (ACR). We also assume that the ACR is not fixed by the intermediate network but by the switches located after the UPC (such that the UPC can keep track of these rate changes). This is a foreseeable situation, at least as long as the VBR and the ABR sources do not cause a heavy congestion condition at the switch in the intermediate network. As a likely consequence of these assumptions, we consider the VBR and the ABR rate changes to be independent.

Finally, we assume a propagation delay equal to zero between the UPC and the ABR source. This is a plausible assumption since it is a fixed delay which should have no influence on the policing function. We also consider that a backward RM cell experiences no delay between the UPC and the ABR source. With these assumptions the only component of the round trip delay

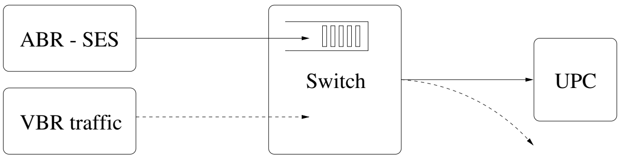


Figure 9.11: System considered in the evaluation.

between the ABR source and the UPC considered in our model is the delay introduced by the switch in the intermediate network.

In the following a detailed description of the analytical model considered for each device is given.

1. **The ABR Source Behavior**

   The ABR traffic that is multiplexed together with the VBR background traffic is described here. A set of N cell rates $r_s,\ s = 1, ..., N$ is associated with the ABR source. The ABR source always transmits at one of these rates for which we have:

   $$\mathrm{MCR} \leq r_1 \leq \ldots \leq r_{N-1} \leq r_N \leq \mathrm{PCR} \tag{9.10}$$

   The parameters MCR and PCR represent the minimum and peak cell rate of the ABR connection. When the ABR source is transmitting at rate $r_i,\ i = 1, ..., N$, a cell is forwarded respectively every $1/r_i,\ i = 1, ..., N$, slots. In our analytical model we need $1/r_i,\ i = 1, ..., N$ to be an integer value.

   The source behavior can be modeled by associating $1/r_i$ states to each possible rate r_i. We denote these states by $(i, j),\ i = 1, ..., N;\ j = 1, ..., 1/r_i$. We say that the ABR source is in state $(i, j),\ j = 1, ..., 1/r_i$, when it is transmitting at rate r_i. As long as the cell rate remains the same, e.g. r_i, the states $(i, 1)$ to $(i, 1/r_i)$ are traversed periodically. i.e. at each slot a transition occurs from (i, j) to $(i, j + 1),\ j = 1, ..., 1/r_i - 1$ (note that in case of $1/r_i = 1$ there is only one of these states, the state $(i, 1)$, and the ABR source remains in it until a transition to a different rate state occurs). A cell is only transmitted in state $(i, 1/r_i)$ and no cell is generated in states $(i, 1)$ to $(i, 1/r_i - 1)$.

   The ABR source considered here does not implement the rescheduling option, therefore, cell transmissions at a new rate are only scheduled at cell emission times (the $(i, 1/r_i),\ i = 1, ..., N$ states in our model). This also implies that when a rate change occurs, e.g. from r_i to $r_{i'}$, there is a transition from state $(i, 1/r_i)$ to state $(i', 1)$. We denote by $P_{\mathrm{ABR}}(i, i'),\ i, i' = 1, ..., N$ the probability that such state transition occurs and assume that they are Markovian.

2. **The Background Traffic**

   We consider as background traffic a VBR source that will be multiplexed together with the ABR source at the switch. This VBR source is modeled by a Markov Chain with M states, each with an associated rate. We say that the source is in state k when it is transmitting at rate $v_k,\ k = 1, ..., M$. These rates obey the following relation:

   $$v_1 < v_2 < \ldots < v_M$$

A cell is generated by the VBR source with probability v_k while being in state k, $k = 1, ..., M$. The VBR source can change its state at the end of each slot. We define $P_{\text{VBR}}(k, k')$ as the probability that a transition occurs from state k to state k', $(k, k' = 1, ..., M)$. Notice that these transitions are independent of the ABR rate changes.

3. **The Switch**

   The switch which precedes the UPC device is used to model the jitter caused by the intermediate network on the ABR traffic. The input of this switch consists of an ABR and a VBR traffic stream generated by the traffic sources described above. As the VBR traffic has full priority over the ABR source and the VBR source never generates more than one cell in a slot, the switch only needs a buffer to store delayed ABR cells. Delayed ABR cells are forwarded by the switch towards the UPC device when there is no VBR cell arrival. If a VBR cell arrives, this cell is forwarded and the ABR cells have to wait.

4. **The UPC device**

   In our model the current state of the UPC device is characterized by its workload U as described in section 9.2. Recall from section 9.2 that the workload U is increased by $\min(I_n^{old}, I_{n+1})$ upon the arrival of cell n, if the cell is accepted, and is decremented by one at the end of each slot. For sake of simplicity we have taken I_n instead of $\min(I_n^{old}, I_{n+1})$ in the analytical model. The validity of this approximation is checked by simulation.

   As described in section 8.2, the expected rate at the interface is computed taking into account the upper and lower bounds of the round trip delay between the UPC and the source, τ_2 and τ_3 respectively. In order to see the influence of these delay bounds we have considered two scenarios. In the first one we model a UPC which immediately applies a rate change when scheduled. This is equivalent to setting $\tau_2 = \tau_3$. Clearly, if there are delayed cells at the switch buffer when a rate decrease occurs, these will be likely considered as non conforming (because an increment higher than the emission interval used by the source will be applied at the UPC).

   In the second scenario we consider a UPC with τ_2 properly set to an upper bound of the round trip delay between the UPC and the source. Such a UPC guarantees that no higher increment than the emission interval used by a "well behaving" source is applied at the UPC. We refer as "well behaving" a source that follows the rate changes conveyed by the backward RM-Cells. Note that this is the kind of source we use in our model. In the following the analytical model we use for these two scenarios is described.

9.4.2 Scenarios Considered in the Evaluation

1. **Scenario with $\tau_2 = \tau_3$**

   In this case we consider a UPC which does not apply a time tolerance to the scheduled rate changes. Since in our model we consider a propagation delay equal to zero, we have $\tau_2 = \tau_3 = 0$. In our model this is equivalent to using the inverse of the rate associated with the ABR state at the time that a cell arrives at the UPC. Clearly this is not necessarily the rate of the ABR source when this cell was generated, as the rate of the ABR source might have changed if the cell was delayed in the switch.

   We recall that as long as the ABR cell rate remains the same, e.g. r_i, it traverses the states $(i, 1), ..., (i, 1/r_i)$ emitting one cell and possibly changing the rate in state $(i, 1/r_i)$. Therefore, we use $1/r_i$, $i = 1, ..., N$ as the increment I_n associated with the states (i, j), $j =$

$1, ..., 1/r_i - 1$. The increment associated with the state $(i, 1/r_i)$ depends on whether or not a rate change occurs. If no rate change occurs the increment $1/r_i$ is used, otherwise we use $1/r_{i'}$ where $r_{i'}$ represents the new rate.

Note that we are making the following approximation. In a real situation with $\tau_2 = \tau_3 = 0$ the UPC would apply I_n immediately after the backward RM-Cell conveying the new rate traversed the UPC. In our model I_n is applied when the ABR source effectively performs the rate change, which happens at the cell emission epoch. This approximation is confronted with simulation results.

2. **Scenario with τ_2 properly tuned**

   In this case the UPC postpones the scheduled rate decreases until after a delay bound $\tau_2 > \tau_3$. Since in our model we consider a propagation delay equal to zero, this implies that during the first τ_2 slots after the scheduling of a rate reduction (from r_i to $r_{i'}$), the UPC will continue to use the smaller increment $1/r_i$.

   We approximate this scenario by flushing the switch buffer each time that a rate reduction occurs. To assess the increments at the UPC we use the same rules as in the previous scenario. Note that by doing this we guarantee that no higher increment than the emission interval used by the source will be applied to a cell arriving at the UPC.

   To be able to solve the analytical model, the flushing is only performed with probability $1 - \alpha$, where α is small, e.g. $\alpha < 10^{-12}$ (see Appendix 9.B).

9.4.3 Performance Analysis

The System Descriptor and the Transition Probability Matrix

The system is observed at the end of each time slot. A Markov Chain is obtained by looking at the following stochastic vector:

$$(Q, U, (i, j), k) \tag{9.11}$$

where Q represents the queue length of the buffer inside the switch, U equals the remaining workload at the UPC, $(i, j), i = 1, ..., N; j = 1, ..., 1/r_i$ is the state of the ABR source and $k, k = 1, ..., M$ the state of the VBR source.

Denote by $P(S, S')$ the one slot transition probability from state $S = (Q, U, (i, j), k)$ to state $S' = (Q', U', (i', j'), k')$. By ordering the states $(Q, U, (i, j), k)$ lexicographically the probabilities $P(S, S')$ define a stochastic transition probability matrix $\mathbf{P}$ with the block structure $\mathbf{P} = (\mathbf{Q}_{m,n})$. The submatrices $\mathbf{Q}_{m,n}$ govern the state $(U, (i, j), k)$ transitions when a queue length change from m to n occurs. Therefore, $\mathbf{Q}_{m,n}$ are square matrices of order equal to $U_{max} + 1$ times the number of ABR states times the number of VBR states, where U_{max} is the maximum workload. Notice that $U_{max} = \tau_1 + 1/r_{min}$, where r_{min} is the minimum of the cell rates considered for the ABR source.

In the Appendixes 9.A and 9.B we describe how to derive the transition probability matrix and how to find the stationary probabilities in each scenario.

The Rejection Probability

Having calculated the stationary probability vector of the process $(Q, U, (i, j), k)$, we denote its components as $\pi(Q, U, (i, j), k)$, i.e. the probability that we are in state $(Q, U, (i, j), k)$. The rejection probability can then be found as:

$$\frac{\left(\sum\limits_{Q\geq 1}\sum\limits_{\substack{U>\\ \tau_1+1}}\sum\limits_{i,j,k}\pi(Q,U,(i,j),k)\,(1-v_k)\right)+\left(\sum\limits_{\substack{U>\\ \tau_1+1}}\sum\limits_{i,k}\pi(0,U,(i,\frac{1}{r_i}),k)\,(1-v_k)\right)}{\left(\sum\limits_{Q\geq 1}\sum\limits_{U,i,j,k}\pi(Q,U,(i,j),k)\ (1-v_k)\right)+\left(\sum\limits_{U,i,k}\pi(0,U,(i,\frac{1}{r_i}),k)\ (1-v_k)\right)} \tag{9.12}$$

The numerator gives us the probability that a cell arrives at the UPC and is rejected. The denominator equals the probability of having an arrival of an ABR cell at the UPC. Note that the first term of the summands considers the case of an ABR cell forwarded from the switch buffer, and the second one considers the case of an ABR cell arrival which finds the switch empty (with no VBR arrival).

9.4.4 Numerical Results

In this section we show the numerical results obtained with the analytical model described in section 9.4.3. In order to validate the analytical model, all the results shown in this section have been verified by simulation.

We have used two different rates for the ABR source. In order to assess $P_{\text{ABR}}(i, i'),\ i, i' = 1, 2$ we have assumed that a Markovian process governs the rate changes. This process alternatively changes between two states, namely $E_i,\ i = 1, 2$. In this model a change into a certain state E_i represents a backward RM-Cell arrival conveying a new rate equal to r_i. Therefore, when the ABR source is in state $(i, 1/r_i),\ i = 1, 2$, a change to state $(i', 1)$ occurs if $E_{i'},\ i' = 1, 2$ is the current state. We have taken the sojourn time in each state $E_i,\ i = 1, 2$, to be the same and equal to p slots. With these assumptions we have:

$$P_{\text{ABR}}(i, i') = \sum_{n=0}^{1/r_i} \binom{1/r_i}{n} 1/p^n\ (1-1/p)^{1/r_i - n} \cdot$$
$$1\big[(i = i' \text{ and } n \text{ is even}) \text{ or } (i \neq i' \text{ and } n \text{ is odd})\big],\ i, i' = 1, 2$$

Note that as long as the inverse of the ABR rates is small compared to p, the mean time between two consecutive rate changes of the ABR source is equal to p. Thus, we will refer to $1/p$ as the ABR rate change frequency.

For the VBR source only one state is possible, and thus one rate v_1. Obviously $P_{\text{VBR}}(k, k') = 1,\ k, k' = 1$. Note that the switch load ρ in this model is approximately given by:

$$\rho = v_1 + \frac{r_1 + r_2}{2}$$

In the following we first investigate the validity of the analytical results by comparing them with the simulation results. Then the figures are analyzed, in order to derive some engineering rules.

Validation

In the simulation we have used the DGCRA given by equations 8.1. Moreover, the approximations made in the analytical model to make it tractable have been removed. The differences are the following:

1. In the analytical model we use I_n instead of $\min(I_n^{old}, I_{n+1})$ (see sections 9.2 and 9.4.1 for the details of this approximation).

2. In the analytical model the increments used by the UPC only change at the ABR source cell emission epochs (see section 9.4.2). In the simulation they are updated at the backward RM-Cell arrival epochs. As explained above, these are given by the state transitions of the Markov process that governs the ABR rate changes.

3. In the analytical model the DGCRA with τ_2 properly tuned is approximated by flushing the queue when a rate reduction occurs (see section 9.4.2). In the simulation, instead, we always use I_n as the emission interval between cell n and $n + 1$.

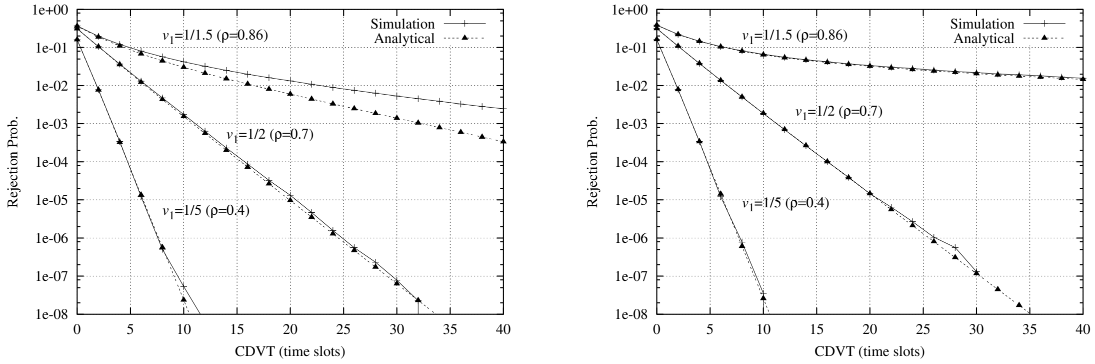
Figures 9.12.A and 9.12.B correspond to the scenario where τ_2 is properly tuned. The figures show a good agreement between the analytical and simulation results. Only when there is a heavy load and the ABR rate change frequency is high, the model yields an underestimated rejection probability.

This can be explained by the following reasoning. If the queue length flushed when there is a rate reduction, e.g. from r_i to $r_{i'}$, is small compared to the cells emitted by the source while the rate was r_i, this approximation will clearly have a small influence on the rejection probability. Therefore, the approximation is worse for high loads and small sojourn times in the states with higher rates.

Figures 9.13.A and 9.13.B show the analytical and simulation results for the scenario where $\tau_2 = \tau_3$. In these figures we can see that the analytical model gives a good approximation for high loads, but it becomes worse when the load decreases. To explain these results we have to take into account the two main reasons that may lead to a cell rejection in this scenario: (i) the jitter of the ABR cell stream and (ii) the usage of an increment higher than the source cell emission interval.

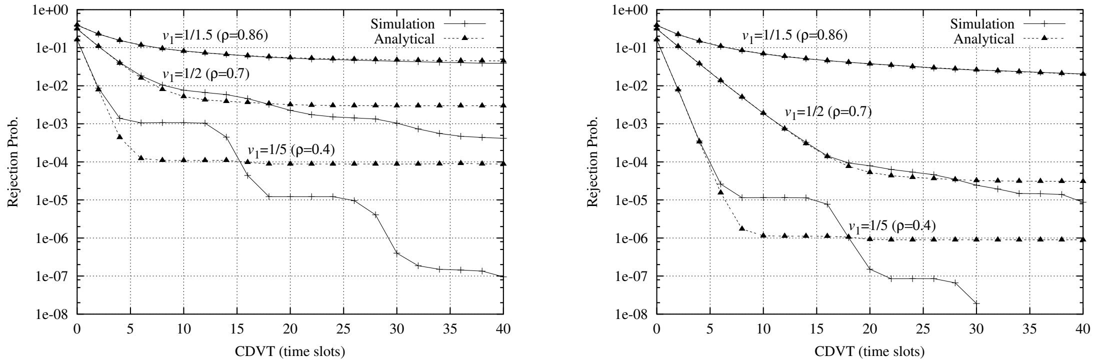
The analytical model is able to capture very well the cells rejected due to condition (i). When the load is high, the condition (i) is predominant and thus the analytical model yields a good approximation. For lower loads, condition (i) is only predominant when the CDVT is close to zero. When the CDVT increases, cell rejection is mainly due to (ii). When this happens the analytical model shows that the rejection probability remains nearly constant. This behavior can be explained by the following reasoning:

Each time a rate reduction occurs, e.g. from r_i to $r_{i'}$, condition (ii) is likely to happen to all cells backlogged in the switch buffer. Therefore, the average workload at the equivalent queue of the DGCRA is approximately increased by $1/r_{i'} - 1/r_i$ times the number of these backlogged cells. Notice that with the same reasoning the average workload can also be reduced at rate increases. However, this workload reduction will not cancel the former increase because the average number of backlogged cells when the rate increases is much smaller (as the source is transmitting at a lower rate) compared to the case of a rate decrease. Furthermore, when a cell rejection occurs the workload is not increased, hence, it is reduced on the average by the



9.12.A: ABR rate change frequency = 1/2009.12.B: ABR rate change frequency = 1/20000

Figure 9.12: Rejection Prob. in the τ_2 properly tuned scenario. ABR rates: $r_1 = 1/3$, $r_2 = 1/15$.



9.13.A: ABR rate change frequency = 1/2009.13.B: ABR rate change frequency = 1/20000

Figure 9.13: Rejection Prob. in the $\tau_2 = \tau_3$ scenario. ABR rates: $r_1 = 1/3$, $r_2 = 1/15$.

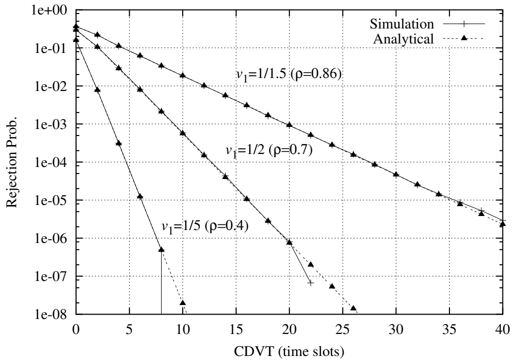


Figure 9.14: Rejection Prob. when there are no ABR rate changes. ABR rate: $r_1 = 1/5$.

emission interval of the rejected cell. In a stationary regime the workload increments must fit the decrements. Therefore, the cell rejection probability is mainly given by the average number of cells to which a wrong increment is applied because condition (ii) occurs, which is independent of the CDVT.

When condition (ii) is predominant, figures 9.13.A and 9.13.B show that the analytical model is initially more optimistic (it yields lower rejection probability), and then the curve obtained by simulation falls down showing a step behavior. The optimistic results are explained because the rate changes are applied later in the analytical model than in the simulation (remember that rate changes occur at the cell emission epochs in the analytical model and at the backward RM-Cell arrivals in the simulation). The steps are caused because $\min(I_n^{old}, I_{n+1})$ is used in the simulation while I_n is used in the analytical model. This makes that each time a rate increase occurs in the simulator, e.g. from r_i to $r_{i'}$, the workload is increased by $1/r_{i'}$ instead of by the emission interval $1/r_i$, which causes an average reduction of $1/r_i - 1/r_{i'}$. Consequently, each time the CDVT added to these reductions compensates the wrong increments mentioned above, there is a reduction on the rejection probability.

The former differences between the simulator (which models the real DGCRA) and the analytical model could be removed. However, this would considerably increase the complexity of the analytical model. For dimensioning purposes the scenario with τ_2 properly tuned would be used, and in this scenario these differences do not have any influence.

Finally, figure 9.14 shows the analytical and simulation results when there are no ABR rate changes. Without rate changes there are no differences between the two scenarios considered in this chapter. Moreover, in this case there are no differences between the simulation and analytical models. This is confirmed by figure 9.14 which shows a perfect agreement of the analytical and simulation results.

Analysis

The figures 9.12-9.14 show the influence of the following items on the rejection probability:

- Tuning of the τ_2 delay bound,
- jitter on the ABR stream (the higher the VBR rate, the higher the jitter),
- difference between the ABR rates,
- frequency of the ABR rate changes.

In figure 9.12 we can see that when τ_2 is properly tuned, the rejection probability decreases exponentially when increasing the CDVT for low to moderated loads. In this case rejection probabilities of 10^{-9} can be reached with CDVT in the order of tens. Moreover, the frequency of the ABR rate changes has little influence on the results.

Figure 9.13 shows that when τ_2 is not properly set, the rejection probability has a major degradation. Moreover, the rejection probability does not decrease exponentially with the CDVT and is much more sensitive to the frequency of the ABR rate changes.

Figure 9.14 shows the rejection probability when the ABR source rate does not change. Note that the average load of the ABR source is maintained constant in all the figures (approximately

```
Initialize UWork = 0, Ik = 1 / ACRk

void WCT_scheduler()
{
    // schedule next WCT cell transmission time

    UWork += Ik - TimeSlot ; // Compute the unfinished
                             // work at the next slot
    if(UWork < CDVT)
        schedule(now() + TimeSlot) ; // schedule at next slot
    else {      // Wait for unfinished work = 0
        schedule(now() + UWork + TimeSlot) ;
        UWork = 0 ;
    }
}
```

Figure 9.15: ABR Source WCT behavior

equal to 0.2). By comparing figure 9.14 with the other figures we can see that the higher the differences between the ABR rates (while maintaining the same load) the higher the rejection probability. The figures also show that this effect increases very rapidly with an increasing overall load.

9.5 ABR Worst Case Traffic

We have seen in section 9.3.1 that a typical ABR stream at the output of a MAN would require a CDVT much higher than at the output of a LAN. Now the question is "how harmful for the network may be an ABR stream with a high CDVT?". To answer this question we have investigated the effect of a source transmitting the worst traffic for the network that would be accepted by a UPC located immediately after the source.

We have assumed that this Worst Case Traffic (WCT) would be given by a source that schedules the cell transmissions according to the algorithm of figure 9.15. Note that such a source computes the unfinished work of the equivalent queue that will be built up at the UPC side (the `UWork` variable in the figure). The source schedules back to back cell transmissions until the unfinished work reaches the CDVT. Then the source delays the cell transmission until the unfinished work goes down to zero and so on.

We have used a network topology consisting of a single switch fed by 10 greedy sources half of which schedule the cell transmissions according to the WCT behavior given by the algorithm of figure 9.15. The source and switch parameters are the same than those used in section 9.3.1. Link delays are 0.5 ms.

Figure 9.16 shows the ACR, the cell rate of one of the WCT sources and the queue length of the switch when the CDVT of the WCT sources have been set to 100 time slots. We can see that the bursts of the WCT sources produce the peaks that are built up in the queue of the switch and the oscillations of the ACR. In figure 9.17 the same measures are shown when the CDVT is adjusted to 1000 time slots. The figure shows that in this case the switch algorithm becomes unstable.

Assuming that an intermediate network could modify an ABR cell stream to behave as the WCT

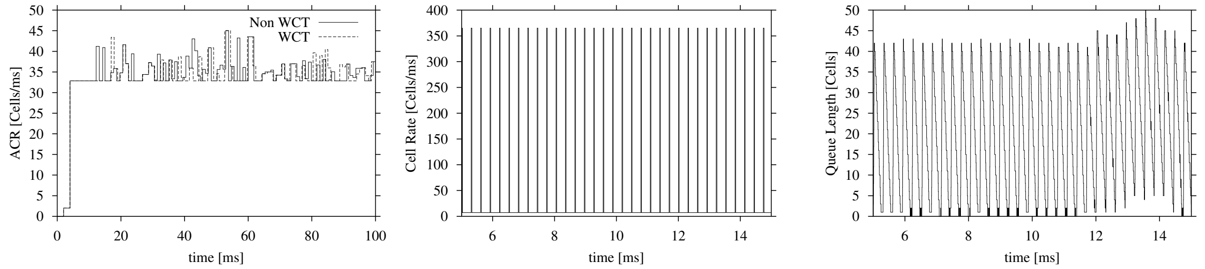


Figure 9.16: Evolution of the ACR, Rate of the ABR-WCT and Queue Length. CDVT = 100 time slots.

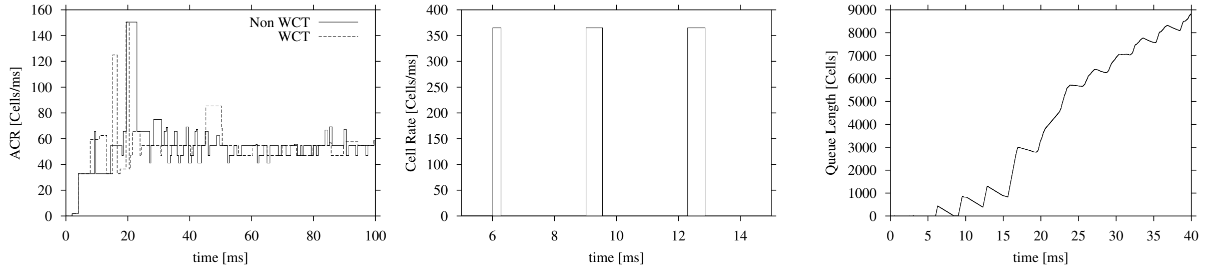


Figure 9.17: Evolution of the ACR, Rate of the ABR-WCT and Queue Length. CDVT = 1000 time slots.

algorithm of figure 9.15 is unrealistic. However, the figures shown in these sections demonstrate that introducing a high CDV in an ABR cell stream may have a negative effect in the stability of the network.

9.6 Conclusions

This chapter investigates the parameter dimensioning in the conformance definition for the ABR Service: the Dynamic Generic Cell Rate Algorithm (DGCRA). An equivalent queuing model of the DGCRA has proposed and solved using two analytical methods: one based on a renewal approximation and another one applying a matrix geometric approach. Furthermore, the negative effect that a high CDVT may have on the network has been studied by means of an "ABR worst case traffic". In the following the conclusions obtained with each of these studies are summarized.

Approach based on a renewal assumption

The renewal assumption allows deriving a simple iterative algorithm to compute the cell rejection probability at the DGCRA. The iterative algorithm use the histograms of the cell inter-arrival and the cell transfer delay variation at the UPC. These histograms can be obtained for a general network, for example, by simulation or measurements.

This approach has been used to analyze two frameworks having different Cell Transfer Delays (CTD). These have been called the LAN and MAN frameworks, the MAN framework having the highest CTD.

The results show that the discrete time analysis is too pessimistic for the LAN but gives a better approximation for the MAN. This is because the renewal assumption applies better in the MAN framework.

Matrix geometric approach

With this approach has been done a finer modeling of the system than with the previous renewal assumption. However, a simpler framework has been considered for sake of applicability of the method.

In the model used in this approach the jitter introduced by a VBR source on an ABR cell stream sharing a common multiplexing stage is considered. The model shows the influence of the following parameters on the cell rejection probability at the UPC: (i) Tuning of the τ_2 delay bound, (ii) jitter on the ABR stream (the higher the VBR rate, the higher the jitter), (iii) the difference between the ABR rates, (iv) the frequency of the ABR rate changes. These are investigated in two scenarios which show the influence of τ_2:

1. Scenario with τ_2 properly tuned:
   - For loads low to moderate, the rejection probability decreases exponentially when increasing the CDVT. In this case, rejection probabilities of 10^{-9} can be reached with CDVT in the order of tens.
   - The higher the differences between the ABR rates (while maintaining the same load) the higher the rejection probability. This effect increases very rapidly with increasing overall loads.
   - The frequency of the ABR rate changes has a minor influence on the rejection probability.
2. Scenario with $\tau_2 = \tau_3$ (the UPC does not apply a time tolerance to the scheduled rate changes):
   - Compared with the former scenario, results show a major degradation of the rejection probability. This probability does not decrease exponentially when increasing the CDVT, but decreases at a slower rate.
   - Frequency and amplitude of the ABR rate changes have a remarkable influence on the rejection probability.

ABR worst case traffic

In order to investigate how harmful for the network may be an ABR stream with a high CDVT, an “ABR worst case traffic” has been defined.

Using this ABR traffic we show that introducing a high CDV in an ABR cell stream may have a negative effect in the stability of the network.

Appendixes

9.A Derivation and Solution of the Transition Probability Matrix in the Scenario with $\tau_2 = \tau_3$

1. **Structure of the Transition Probability Matrix**

   Since the ABR source can only emit one cell at each slot, $P(S, S')$ elements with queue length increments or decrements higher than one (i.e. $|Q' - Q| > 1$) are zero. Moreover, the transitions are exactly the same for all values with $Q \geq 1$. Therefore, we define: $\mathbf{B}_0 = \mathbf{Q}_{0,0}$, $\mathbf{A}_0 = \mathbf{Q}_{n+1,n},\ n > 0$, $\mathbf{A}_1 = \mathbf{Q}_{n,n},\ n > 0$ and $\mathbf{A}_2 = \mathbf{Q}_{n,n+1},\ n \geq 0$. This yields the following structure for the matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{pmatrix} \mathbf{B}_0 & \mathbf{A}_2 & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \dots \\ \dots & \dots & \dots & \dots & \ddots \end{pmatrix} \tag{9.13}$$

   The following partitioned solution $(\pi_0, \pi_1, ...)$ of the stationary probabilities exists for these type of processes [60] ($\pi_i,\ i = 0, 1, ...$ are vectors of length equal to the order of the blocks of the matrix $\mathbf{P}$):

$$\begin{array}{ll} \pi_k = \pi_0\ \mathbf{R}^k & ,k \geq 1 \\ \pi_0 = \pi_0\left[\mathbf{B}_0 + \mathbf{R}\mathbf{A}_0\right] & \\ \pi_0\left(\mathbf{I} - \mathbf{R}\right)^{-1}\mathbf{e} = 1 & \end{array} \tag{9.14}$$

   where $\mathbf{R}$ has the same order as $\mathbf{A}_i$. $\mathbf{I}$ is the unity matrix with the same order and $\mathbf{e}$ is an all ones vector with corresponding length. To find $\mathbf{R}$ we use the logarithmic reduction algorithm of Latouche and Ramaswami [51].

2. **Derivation of the blocks of the Matrix P**

   First we make some remarks concerning the notation: by $\mathbf{A}(S, S')$ we denote the element (S, S') of the matrix $\mathbf{A}$. Remember from section 9.4.3 that the element (S, S') of a block of $\mathbf{P}$ represents the one slot transition probability from state $(U, (i, j), k)$ to state $(U', (i', j'), k')$, where U equals the remaining workload at the UPC, (i, j) is the state of the ABR source and k the state of the VBR source. By $[x]^+$ we shall refer to $\max(x, 0)$. Finally, we shall use the indicator function $1\left[\mathit{condition}\right]$ equal to 1 if *condition* is true and 0 otherwise.

   Derivation of $\mathbf{A}_2$: In this case the queue length at the switch is increased by one, i.e. $Q' = Q + 1$. This can only happen if a cell is emitted by the ABR source (and thus $j = 1/r_i$) and a cell is emitted by the VBR source. Remember from section 9.4.1 that the ABR source changes to state $j' = 1$ after a cell emission. Since there is no cell arrival at the UPC the workload is decreased by 1, and thus $U' = [U - 1]^+$. Therefore, we have:

$$\mathbf{A}_2(S, S') = P_{\text{ABR}}(i, i') \cdot P_{\text{VBR}}(k, k') \cdot v_k \cdot 1\left[j = 1/r_i \text{ and } j' = 1 \text{ and } U' = [U - 1]^+\right]$$

   The following relations are obtained with similar reasoning.

Derivation of $\mathbf{A}_0$: $(Q' = Q - 1,\ Q > 0)$

$$\begin{aligned}\mathbf{A}_0(S,S') = P_{\mathrm{VBR}}(k,k') \cdot (1-v_k) \cdot 1\big[i = i' \text{ and } j < 1/r_i \text{ and } j' = j+1 \text{ and}\\ \big((U-1 \le \tau_1 \text{ and } U' = [U-1]^+ + \frac{1}{r_{i'}}) \text{ or } (U-1 > \tau_1 \text{ and } U' = [U-1]^+)\big)\big]\end{aligned}$$

Derivation of $\mathbf{A}_1$: $(Q' = Q,\ Q > 0)$

$$\begin{aligned}\mathbf{A}_1(S,S') = P_{\mathrm{VBR}}(k,k') \cdot \Big(P_{\mathrm{ABR}}(i,i') \cdot (1-v_k) \cdot 1\big[j = 1/r_i \text{ and } j' = 1 \text{ and}\\ \big((U-1 \le \tau_1 \text{ and } U' = [U-1]^+ + \frac{1}{r_{i'}}) \text{ or } (U-1 > \tau_1 \text{ and } U' = [U-1]^+)\big)\big] +\\ v_k \cdot 1\big[i = i' \text{ and } j < 1/r_i \text{ and } j' = j+1 \text{ and } U' = [U-1]^+\big]\Big)\end{aligned}$$

Derivation of $\mathbf{B}_0$: $(Q' = 0,\ Q = 0)$

$$\begin{aligned}\mathbf{B}_0(S,S') = \mathbf{A}_1(S,S') + P_{\mathrm{VBR}}(k,k') \cdot (1-v_k) \cdot\\ 1\big[i = i' \text{ and } j < 1/r_i \text{ and } j' = j+1 \text{ and } U' = [U-1]^+\big]\end{aligned}$$

9.B Derivation and Solution of the Transition Probability Matrix in the Scenario with τ_2 Properly Tuned

1. **Structure of the Transition Probability Matrix**

   We approximate this scenario by flushing the switch buffer each time that a rate reduction occurs, i.e $r_i > r_{i'}$ (see section 9.4.2). For the reasons explained below, the flushing is only performed with probability $1-\alpha$. A transition from $Q \ge 0$ to $Q' = 0$ occurs when the queue is flushed, therefore, we obtain the following structure for the transition probability matrix $\mathbf{P}^{(p)}$ (by the superscript $^{(p)}$ we shall distinguish the matrices derived in this scenario from those used in section 9.4.2):

$$\mathbf{P}^{(p)} = \begin{pmatrix} \mathbf{B}_0^{(p)} & \mathbf{A}_2^{(p)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}_0^{(p)} + \mathbf{A}_3^{(p)} & \mathbf{A}_1^{(p)} & \mathbf{A}_2^{(p)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}_3^{(p)} & \mathbf{A}_0^{(p)} & \mathbf{A}_1^{(p)} & \mathbf{A}_2^{(p)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}_3^{(p)} & \mathbf{0} & \mathbf{A}_0^{(p)} & \mathbf{A}_1^{(p)} & \mathbf{A}_2^{(p)} & \mathbf{0} & \cdots \\ \mathbf{A}_3^{(p)} & \mathbf{0} & \mathbf{0} & \mathbf{A}_0^{(p)} & \mathbf{A}_1^{(p)} & \mathbf{A}_2^{(p)} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} \tag{9.15}$$

2. **Derivation of the blocks of the Matrix $\mathbf{P}^{(p)}$**

   The submatrices are given by:

$$\begin{aligned} &\mathbf{A}_0^{(p)} = \mathbf{A}_0 \qquad &&\mathbf{A}_1^{(p)} = \mathbf{A}'_1 \qquad &&\mathbf{A}_2^{(p)} = \mathbf{A}'_2\\ &\mathbf{A}_3^{(p)} = \mathbf{A}''_1 + \mathbf{A}''_2 \qquad &&\mathbf{B}_0^{(p)} = \mathbf{B}_0 + \mathbf{A}''_2 \end{aligned}$$

   where the matrices $\mathbf{A}_0$ and $\mathbf{B}_0$ are obtained as in section 9.A. The matrices $\mathbf{A}'_1$ and $\mathbf{A}'_2$ are respectively obtained replacing $P_{\mathrm{ABR}}(i,i')$ by $P_{\mathrm{ABR}}(i,i') \cdot \big(1\big[r_i \le r_{i'}\big] + \alpha \cdot 1\big[r_i > r_{i'}\big]\big)$ in the relations given for $\mathbf{A}_1$ and $\mathbf{A}_2$ in section 9.A. Similarly, the matrices $\mathbf{A}''_1$ and $\mathbf{A}''_2$

are obtained replacing $P_{\text{ABR}}(i,i')$ by $P_{\text{ABR}}(i,i') \cdot (1-\alpha) \cdot 1\big[r_i > r_{i'}\big]$ in the relations of $\mathbf{A}_1$ and $\mathbf{A}_2$.

To solve the former matrix $\mathbf{P}^{(p)}$, the matrix $\mathbf{A}_0^{(p)} + \mathbf{A}_1^{(p)} + \mathbf{A}_2^{(p)}$ must be irreducible [59]. By choosing $\alpha > 0$, this condition is fulfilled. $\mathbf{P}^{(p)}$ is then solved using equations (9.14) except that the second equation has to be replaced by:

$$\pi_0 = \pi_0 \left[\mathbf{B}_0^{(p)} + \mathbf{R}\ \mathbf{A}_0^{(p)} + \big((\mathbf{I}-\mathbf{R})^{-1} - \mathbf{I}\big)\ \mathbf{A}_3^{(p)}\right] \tag{9.16}$$

The matrix $\mathbf{R}$ still obeys a similar equation as before and therefore can be found using the L-R algorithm [51].

Chapter 10

Charging of ABR

10.1 Introduction

The charging scheme that will be applied to ABR has been not specified by the standardization bodies. Pricing however may be an essential condition for the users when submitting traffic. Consider for example a charging scheme based on the allocated bandwidth. With such scheme an ABR source may decide to constrain the bandwidth demand to a lower value than the offered by the fair division of the free capacity. This would allow a redistribution of the available bandwidth giving more to those users that place the highest value on it. Therefore, by means of prices the network can indirectly modify the source rates, and thus, prices introduce a different meaning to the fairness concept. In [45] an analytical approach can be found which takes into account rates and prices, and shows their influence on the fairness criteria.

Internet gives a practical example. By now, Internet offers a unique “best effort” service. Typically, connections of large organizations are leased lines charged with an annual fee based on the bandwidth of the access link, and give unlimited access. The absence of a usage charge has undoubtedly stimulated the usage and development of Internet. This pricing scheme, however, provides no incentive to efficiently adapt the user needs to the network resources. Due to the enormous growth of new users and bandwidth demand, many proposals have been made justifying the adoption of a multiservice and usage-based pricing model (see for example [54, 24, 72]).

An interesting usage-based pricing proposal is made by MaKie-Mason and Varian [53]. This model, referred as “smart market”, consists of the users setting a “bid” on each packet for the maximum willingness-to-pay. In case of congestion the routers discard the packets with lowest bids. The discarded packet with the highest bid is the price charged to all users. This pricing model has many desirable features. It is easy to understand and predictable by the user (the bid fixes the maximum charge). Moreover, it can be shown that the best choice for the users is setting the true value for the bids.

An usage-based pricing is also desirable for an ATM network, the question is how the usage-based scheme should be selected? A fundamental economic principle is that prices should reflect cost. For sources with guaranteed bandwidth as CBR and VBR, the allocated bandwidth or the generalized concept of the effective bandwidth, and the duration of the connection is a good characterization of their resource usage. Charging of ABR may be more complex because it is the responsibility of the network to fairly divide the network resources among the sources, but the network may not have an a priory knowledge of the users appraisal of such resources.

Songhurst and Kelly [73] propose a unified pricing model for VBR and ABR services which consists of fixing the charging parameters at the connection set up, and computing the connection charge by multiplying these parameters by the duration and number of transmitted cells. For the VBR service the charging parameters are computed based on the effective bandwidth, while for ABR the charging parameters are computed based on the requested MCR. The authors assume that the network divides the free bandwidth proportionally to the MCR. MCR is therefore the mean for the users to communicate their preferences to the network.

We refer the previous approach as a "static" charging model because the charging parameters are established at the connection set up and do not change afterwards. Another approach consists of dynamically varying the charging parameters in order to adapt the source demand to the network capacity. L. Murphy and J. Murphy first proposed this pricing framework for ATM, see for example [58]. In the proposed framework, the users are assumed to have a benefit function which completely determines their bandwidth request given a price per unit of bandwidth. The network operator dynamically adjusts the prices based on monitored network conditions and users reduce or increase their bandwidth demand accordingly. However, the approach is described generically and is intended to be applied to the full range of service types. Based on the same ideas, a dynamic charging for the ABR service has been proposed by Courcoubetis et al. [25]. The method consists of using the forward RM-cells of the ABR service to convey the source demand to the switches, and the backward RM-cells to convey the prices to the sources. Both demand and prices are adjusted by an iterative algorithm which in equilibrium satisfies the source demand and maximizes the network revenue.

The chapter is organized as follows. Section 10.2 analyzes the static scheme suggested by Songhurst and Kelly in [73]. In this scheme, charges are computed based on the duration of the connection. We propose a new alternative based on the volume of transmitted cells. Section 10.3 analyzes the dynamic charging model proposed by Courcoubetis et al. in [25]. In this charging scheme the sources convey their demand to the switches, which use this information to compute the prices. We propose an alternative scheme which uses the switch loads and show analytically the evolution of the prices. We describe how to integrate this scheme in an ABR switch and analyze it by simulation. Finally, a numerical comparison of the different schemes is performed and give some concluding remarks in sections 10.4 and 10.5 respectively.

10.2 Static Pricing Schemes

Songhurst and Kelly [73] propose a pricing model where the charge of a connection is given by the expression:

$$\text{Total Charge of a connection} = a(x)T + b(x)V + c(x) \tag{10.1}$$

where T is the duration of the connection, V the volume (number of cells) submitted by the connection and x is the tariff choice which includes the service class, the traffic contract parameters and others. $c(x)$ is a fixed subscription fee. Based on the effective bandwidth, for VBR connections Songhurst and Kelly [73] give an expression for $a(x)$ and $b(x)$ which minimizes the expected charge given the expected mean and peak rate of the connection. Therefore, the scheme encourages the user to give an accurate value of the traffic contract parameters.

Songhurst and Kelly [73] propose to use expression (10.1) also for charging ABR sources. In this case $a(x)$ is assessed proportionally to the MCR, i.e. $a(x) = \gamma \cdot \text{MCR}$, and $b(x)$ is assumed to be much lower than γ or even zero. Note that if the connection transmits always at a rate $r \geq$ MCR, then MCR $\cdot\, T$ is the volume of traffic submitted at the guaranteed MCR. Therefore, it can be interpreted that the traffic transmitted above the MCR is charged at a much cheaper price. In order for the sources to select an MCR according to their bandwidth appraisal, the authors assume that the network divides the free bandwidth proportionally to the MCR.

10.2.1 A Static Pricing Model with Customizable per cell Tariffs

The static pricing scheme described in the previous section has the drawback that many data sources may not be able to choose an adequate MCR due to their bursty nature. For such sources a pricing model which charges the transmitted volume rather than the duration of the connections would fit better.

The model we suggest would consist of several prices per cell tariffs p_i for the ABR service. Users would select a tariff p_i at the connection set up in order to charge the cells transmitted at the shared bandwidth. We further assume that the network allocation algorithm should divide the free bandwidth proportionally to the chosen tariff p_i. Of course, some sources may need a guaranteed MCR, which is likely to be charged based on the duration of the connection. Thus, we propose the following charging equation:

$$\text{Total Charge of a connection} = \gamma \cdot \text{MCR} \cdot T + p_i \cdot \max\left\{V - \text{MCR} \cdot T, 0\right\} \tag{10.2}$$

Note that if the connection transmits always at a rate $r \geq$ MCR, then MCR $\cdot\, T$ is the volume of traffic submitted at the guaranteed MCR. Therefore, the first term of the right side of equation (10.2) charges this traffic at γ [unit of price/ cell]. The second part of equation (10.2) is intended to charge the volume of traffic transmitted above the MCR at the price p_i chosen by the user, $p_i < \gamma$. Clearly, when the source rate $r \geq$ MCR, the cells given by $\max\left\{V - \text{MCR} \cdot T, 0\right\}$ are the cells transmitted above the MCR. This is not a drawback, however, because it would penalize users which choose a guaranteed MCR higher than their needs.

10.3 Dynamic Pricing Schemes

Based on the principle of social welfare optimization [52], a dynamic charging for the ABR Service has been proposed by Courcoubetis et al. [25]. The method consists of using the forward RM-cells of the ABR service to convey the source demand to the switches, and the backward RM-cells to convey the prices to the sources. Both demand and prices are adjusted by an iterative algorithm which in equilibrium satisfies the source demand and maximizes the network revenue (maximum social welfare). In the following this method is briefly described.

Let C_l be the capacity available for ABR traffic traversing link l and a_l the price per transmitted cell charged to each connection traversing link l. Denote R_c the route of connection c and w_c the charge of connection c. Clearly $w_c = \sum_{l \in R_c} a_l$. Let assume that each connection c has a bandwidth demand equal to $D_c(w_c)$. Finally denote x_c the actual rate of connection c.

In [25] the authors show that the maximum social welfare is satisfied by the following relations:

$$x_c = D_c(w_c), \quad \text{for all } c \tag{10.3}$$

$$\sum_{c:l\in R_c} x_c \le C_l, \quad \text{for all } l \tag{10.4}$$

$$a_l(C_l - \sum_{c:l\in R_c} x_c) = 0, \quad \text{for all } l \tag{10.5}$$

Equation (10.3) says that the social welfare is maximized when the connection rates equal their demands. Note that equations (10.4) and (10.5) imply that $a_l = 0$ for the non congested links ($\sum x_c < C_l$), and $a_l \neq 0$ for the congested links ($\sum x_c = C_l$). In [25] an iterative scheme which converges to this equilibrium is proposed. The algorithm consists of the links updating the price a_l^n at fixed intervals n of duration δ (referred as charging intervals). Prices at interval n are decreased or increased if $\sum D_c^{n-1} < C_l$ or $\sum D_c^{n-1} > C_l$ respectively.

In order to integrate this charging scheme in the ABR flow control, the authors in [25] propose to add two new fields to the RM-cells: a request bandwidth (RB) field and a price per unit of bandwidth (PB) field. Based on the prices w_c, the sources set the RB field with the demand function $D_c(w_c)$. These values are used by the switches to compute the prices a_l^n applying an iterative algorithm. The switches increase the PB field of all the backward RM cells by a_l^n. Thus, when the backward RM cell arrives to the source it conveys $\sum_{l\in R_c} a_l$ in the PB field. Note that a billing unit located at the network edge keeping track of the PB field of the backward RM-cells should be needed to compute charges.

10.3.1 A Dynamic Pricing Model Based on the Switch Loads

A problem of the pricing model previously described is that users do not have an incentive to specify their true demand. Users could set a misleading demand value in order to modify the prices (for example, the demand field could be set always to zero). To solve this problem we propose to use the offered load ($\sum_{c:l\in R_c} x_c^{n-1}$) instead of the source demand ($\sum_{c:l\in R_c} D_c^{n-1}$) to decide when to increase or decrease prices. Obviously, the RB field of the RM-cells conveying the request bandwidth of the sources is not needed anymore with this algorithm. Just the PB field would be used as described in the previous section.

Assume that the control algorithm of the switch adjusts the source rates such that the offered load of link l converges to a certain target cell rate TCR_l. We define the overload as $O_l^{n-1} = \sum_{c:l\in R_c} x_c^{n-1}/\text{TCR}_l$. Based on the overload we propose the following algorithm to compute the prices:

$$a_l^n = \begin{cases} \max\left\{(1 + h\,\text{sgn}\left[O_l^{n-1} - \alpha\right])\,a_l^{n-1}, 0\right\}, & \text{if} \quad a_l^{n-1} \neq 0 \\ a_0, & \text{if} \quad a_l^{n-1} = 0 \end{cases} \tag{10.6}$$

where $\text{sgn}[x] = \begin{cases} +1, & \text{if } x \ge 0 \\ -1, & \text{if } x < 0 \end{cases}$ and $\alpha \lesssim 1$ is a constant parameter. Note that now prices are adjusted such that the link load converges to $\alpha \cdot \text{TCR}_l$. In the following the algorithm (10.6) is analyzed.

For simplicity, assume a network with only one switch and all the sources located at the same distance from the switch. Let refer τ as the round trip delay from the sources to the switch.

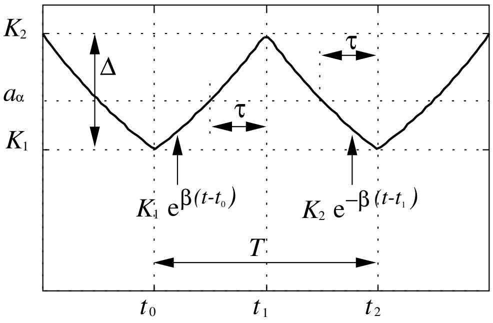


Figure 10.1: Evolution of the prices applying algorithm (10.6).

First note that in equilibrium, if $a_l^n \geq 0$, equation 10.6 may be approximated by:

$$a(t) = a(t{-}\delta) + \text{sgn}[O(t{-}\delta) - \alpha]\ h\, a(t{-}\delta) \tag{10.7}$$

With the approximation $\frac{d}{dt}w(t) \approx \frac{w(t)-w(t-\delta)}{\delta}$ equation (10.7) becomes:

$$\frac{d}{dt}a(t) = \frac{h}{\delta}\text{sgn}[O(t) - \alpha]\ a(t) \tag{10.8}$$

The solutions of equation (10.8) are of the form ($\beta = h/\delta$):

$$a(t) = \begin{cases} K_1\, e^{\beta t}, & \text{if } O(t) \geq \alpha \\ K_2\, e^{-\beta t}, & \text{if } O(t) < \alpha \end{cases} \tag{10.9}$$

Assume that at time t_0^+, $O(t_0^+) > \alpha$ and equation $K_1\, e^{\beta(t-t_0)}$ applies (cfr. figure 10.1). The price increases until the time instant t_1, at which the overload at the switch falls below α. In equilibrium the source rate is equal to the demand (equation (10.3)). We have $O(t) = \sum D_c(a(t-\tau))/\text{TCR}$. Let a_α be the price at which $\sum D_c(a_\alpha)/\text{TCR} = \alpha$. At time $t_1 - \tau$ the price $a(t)$ reaches a_α at the switch. This value is received by the sources at $t_1 - \tau/2$. The corresponding rate reduction is received by the switch at time t_1. We have the following relations (cfr. figure 10.1): $K_2 = K_1\, e^{\beta\,(t_1-t_0)}$, $K_1 = K_2\, e^{-\beta\,(t_2-t_1)}$, $K_2\, e^{-\beta\,(t_2-\tau-t_1)} = a_\alpha$, $K_1\, e^{\beta\,(t_1-\tau-t_0)} = a_\alpha$, which imply: $t_1 - t_0 = t_2 - t_1 = 2\,\tau$, $K_1 = a_\alpha\, e^{-\beta\,\tau}$, $K_2 = a_\alpha\, e^{\beta\,\tau}$, and thus, the period ($T$) and the amplitude ($\Delta$) of the oscillation are given by:

$$T = 4\,\tau \tag{10.10}$$

$$\Delta = a_\alpha\,(e^{\beta\,\tau} - e^{-\beta\,\tau}) \tag{10.11}$$

Note that the previous algorithm is always stable and easy to tune. The only parameter h could be fixed by the following reasoning. Assume that the design criteria is that the oscillation relative to the equilibrium price has to be lower than a certain bound B_{max} (i.e. $\Delta/a_\alpha \leq B_{max}$). If the amplitude of the oscillation is small, from equation (10.11) we have $\Delta \approx a_\alpha\, 2\,\beta\,\tau$. Remember that $\beta = h/\delta$, therefore, the previous bound is equivalent to:

$$h \leq \frac{\delta}{2\,\tau}B_{max} \tag{10.12}$$

Given the maximum round trip delay from a source to the switch, and a desired B_{max}, equation (10.12) could be used to assess the value of h. Note however that the time constant of the algorithm is $1/\beta = \delta/h$. Therefore, reducing the value of h reduces also the convergence speed of the algorithm.

10.3.2 Integration of the Dynamic Pricing in the ERICA Switch

This section describes how the Dynamic Pricing scheme introduced in the previous section could be integrated in the well known ERICA switch. Remember from section 5.2.3 that ERICA is an Explicit Rate (ER) switch, i.e. it divides the available rate among the contending sources and conveys the fair rate to the sources by means of the ER field of the RM cells. In order to compute the fair rates, the ERICA switch computes the switch overload at each measuring interval given by N cell arrivals.

The switch could therefore apply algorithm (10.6) to compute the prices at each measuring interval, just after computing the overload. Then the PB field of the backward RM-cells could be increased with the computed price at the same time as the ER field is set to the fair rate. In this case, the charging interval δ described in section 10.3 would correspond to the measuring interval. Assuming that cells arrive at target cell rate (TCR), this interval would be given by $\delta = N/\mathrm{TCR}$. Substituting in (10.12), the parameter h of the charging algorithm would be given by:

$$h \leq \frac{N}{2\,\mathrm{TCR}\,\tau} B_{max} \tag{10.13}$$

10.3.3 Simulation Analysis

This section gives a pictorial view of the source rates and prices that would be obtained applying the dynamic model previously described. The network topology of figure 10.2 is considered. All links (arrows in the figure) have a propagation delay of 0.5 ms (equivalent to a distance of 100 km for a propagation delay of 5 μs/km). The circle represents a FIFO queue which feeds a common link with a capacity of 365 cells/ms and implements the ERICA flow control algorithm with parameters TCR $= 0.9 \cdot$ Link Rate and Counting Interval $= 100$ cells. The switch initializes the pricing parameter $a_l^0 = 0$ and implements the algorithm previously described to update its value. To adjust the parameter h of the algorithm we have applied the formula (10.13) choosing $B_{max} = 0.02$. Propagation delay is 0.5 ms, thus $\tau = 1$ ms and substituting we have $h = 0.03$. The parameter α of the algorithm has been set to 0.95 and $a_0 = 0.01$.

Sources are assumed to have always cells to transmit and adjust the transmission rate to $\min\{D_c(w),\ \mathrm{ER}\}$ at each backward RM-cell arrival (greedy behavior), where w is the price per unit of bandwidth and ER the explicit rate carried by the RM-cell. Source parameters are RIF $= 1$, Nrm $= 32$, ICR $= 2$ cells/ms and PCR $=$ Link Rate. We choose the source demand assuming that the user criterion is to fix a maximum charge per unit of time p_c. Remember that

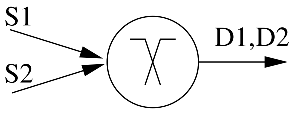


Figure 10.2: Network topology

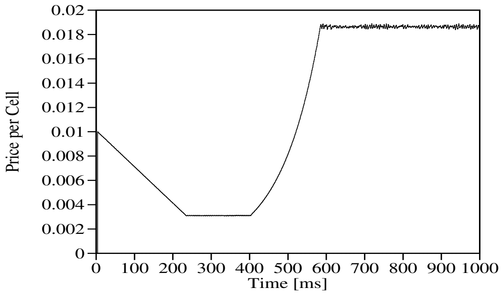


Figure 10.3: Prices computed at the switch.

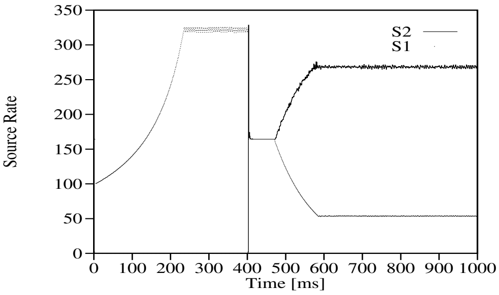


Figure 10.4: Sources rates.

being w the price per transmitted cell posted by the network and D the source rate, the charge after a time period T becomes: charge $= w \cdot D \cdot T$. Therefore, to upper bound the charge per unit of time by p_c (i.e. charge$/T \leq p_c$), the source demand has to be $D \leq p_c/w$. The source is upper bounded by the PCR, thus such a source demand function is given by:

$$D_c(w) = \min\{p_c/w,\ \mathrm{PCR}\} \tag{10.14}$$

We suppose that source S2 is ready to pay 5 times more than source S1, and thus we have set $p_c = 5$ for source S2 and $p_c = 1$ for source S1.

Figures 10.3 and 10.4 show the evolution of the prices computed at the switch and the source rates. Source S2 is staggered 400 ms from source S1. Thus, initially the price is stabilized such that the source rate of S1 reaches the TCR of the switch. When the source S2 becomes active the prices increases until source S2 rate is 5 times the rate of S1, according to the source demands.

10.4 Numerical Comparison

This section performs a numerical comparison of the pricing schemes described in this chapter, namely, the Static Pricing model based on the Duration of the connection (SPD), the one based on per cell tariffs (SPC) described in section 10.2, and the dynamic pricing model (DP) based on the switch loads described in section 10.3.

Again, the network topology of figure 10.2 has been assumed, but now the switch is fed by a greedy (S1), an ON-OFF source (O1), and a background of 5 greedy sources (S2). The links have a propagation delay of 0.05 ms. In order to select the source parameters we assume that the user criterion is to fix a maximum charge per unit of time α. We also assume that the sources are not able to predict their traffic pattern. Finally, sources S1 and O1 are supposed ready to pay double in order to obtain a higher bandwidth, thus sources S1 and O1 choose $\alpha = 2$ and sources S2 choose $\alpha = 1$.

In the SPD model, the price per unit of time is proportional to the MCR. We thus take, without loss of generality, MCR $= \alpha$ and the charging formula will be given by equation (10.1) with $a(x) = \alpha$, $b(x) = 0$, $c(x) = 0$. Remember that the switch is assumed to divide the available bandwidth proportionally to the MCR, therefore, the switch applies a weight $= \alpha$ in this scheme. In the DP scheme the above criterion implies a demand function given by equation (10.14) with $p_c = \alpha$. The pricing parameter of the switch has been adjust to $h = 0.005$. Finally, in the SPC scheme we assume that the user selects MCR $= 0$ and upper bounds the charge per time unit by choosing a price per cell (and thus a switch weight) equal to $p_i = \alpha/\mathrm{PCR}$ (cfr. equation (10.2)).

	Pricing based on time (SPD)			Pricing based on per cell tariff (SPC)			Dynamic pricing (DP)		
	O1	S1	S2	O1	S1	S2	O1	S1	S2
PCR	100	100	100	100	100	100	100	100	100
MCR	2	2	1	0	0	0	0	0	0
Weight	2	2	1	2/PCR	2/PCR	1/PCR	-	-	-
Demand	-	-	-	-	-	-	$2/w$	$2/w$	$1/w$

Table 10.1: Source parameters.

Type of pricing	Source	Rate [Cells/ ms]	Duration [ms]	Usage	Total Charge	Charge per Cell	Charge per Time	Expected Charge per Time
Based on time (SPD)	O1	71.9	729	0.177	1459	0.157	2	0.354
	S1	91.0	1000	1	2000	0.022	2	2
	S2	45.5	1000	1	1000	0.022	1	1
Dynamic pricing (DP)	O1	63.2	747	0.197	249	0.027	0.33	0.394
	S1	86.5	1000	1	1925	0.022	1.92	2
	S2	44.8	1000	1	988	0.022	0.98	1
Per cell tariff (SPC)	O1	71.9	729	0.177	186	0.02	0.255	0.354
	S1	90.6	1000	1	1820	0.02	1.82	2
	S2	45.4	1000	1	454	0.01	0.454	1

Table 10.2: Measures.

Table 10.1 summarizes the source parameters corresponding to each pricing scheme. In the simulation we assume that the greedy sources become active at time $t = 0$ and remain active until the end of the simulation at $t = 1000$ ms. The ON-OFF source becomes active at $t = 100$ ms and transmits three bursts of 3000 cells spaced 300 ms (after the end of the previous burst). Bursts are transmitted with a greedy behavior and the source finalizes the transmission after the end of the third burst. The switch and source parameters not mentioned here are the same that those applied in section 10.3.3. In the simulations carried out with the static pricing schemes, a weighted switch applying the weights shown in table 10.1 was used. This weighted switch is explained in appendix 10.A. With the dynamic pricing the ERICA switch was used (the ERICA switch is explained in section 5.2.3). We have to use the ERICA switch because with the weighted switch, the switch and the charging algorithms interfere and the system exhibits strong oscillations.

Table 10.2 shows the measures taken during the simulation applying the three pricing schemes. These measures are the following:

- The average cell rate,
- the duration of the connection,
- the usage. We define this parameter as the ratio of the active period of the source to the duration of the connection. This value is 1 for the greedy sources and lower for the ON-OFF source,
- the charges,

- the expected charge per unit of time which is the value α chosen by the source multiplied by the usage. In our framework, this value could be taken as the goal of the charging scheme.

In the case of SPD, the user exactly pays the selected charge per time, regardless of the active or idle state of the source. This severely penalizes the bursty source O1, widely exceeding the expected charge per time. In this sense, the DP scheme yields a more fair charge, and all the O1, S1 and S2 sources nearly achieve the expected charge per unit of time. Finally, in the SPC scheme, the charge per cell is fixed. Remember that this has been chosen to be $p_i = \alpha/\text{PCR}$, thus the difference between the measured and expected charge per time will be as high as the difference between the PCR and the average rate. However, this pricing scheme does not penalize the bursty source O1 and therefore is a better solution for charging bursty sources than the other static scheme while being less complex than the dynamic one.

10.5 Conclusions

Charging of ABR is an open issue. This chapter analyzes some existing proposals and suggest new alternatives. Charging approaches have been classified as “static” when prices are established at the connection set up and do not change afterwards, and “dynamic” otherwise.

A numerical comparison is performed of two static schemes, which respectively charges the duration (SPD) and the volume (SPC) of the connection, and a dynamic scheme (DP). The comparison shows that the SPD scheme severely penalizes bursty sources. The DP scheme on the contrary is able to adapt prices to source usage and appraisal of network resources, giving the best-expected charge. Finally, the SPC gives an intermediate performance.

Static charging schemes, however, have the advantage of simplicity, although switches are required to divide the available bandwidth according to a weight related to the applied tariff (weighted allocation). Instead, dynamic charging schemes are more complex since switches have to update the prices continuously according to the link occupancy. However, weighted allocation is not needed in the dynamic schemes.

Appendixes

10.A A Weighted Switch Algorithm

The pricing schemes described in section 10.2 assume that the switch algorithm allocates the available bandwidth among the contending sources proportionally to a weight. However, the switch algorithms proposed up to now, have mostly considered a max-min allocation.

In this section a weighted switch algorithm is described. The algorithm is similar to the one described in [56].

Let be B the available bandwidth to be divided among the ABR contending sources, and w_j the weight to be applied in the bandwidth allocation to the source j. The goal of the weighted algorithm is to allocate a rate B_j given by:

$$B_j = MCR_j + \frac{(B - \sum_i MCR_i)}{\sum_i w_i}\, w_j \tag{10.15}$$

To achieve this goal we use the algorithm shown in the pseudo-code of figure 10.5. In order to apply equation (10.15), the algorithm estimates $(B - \sum_i MCR_i)/\sum_i w_i$, referred as `FairShare` in the figure. From equation (10.15) we have:

$$\texttt{FairShare} = (B_j - MCR_j)/w_j \tag{10.16}$$

The algorithm applies equation (10.16) to the allocated rates B_j at each measuring interval, and takes the maximum, referred as `MaxAllocated` in the figure (sentences 7∼9). Note that the measuring interval is given by a cell counting up to `MIC` cells (cfr. sentence 11). In order to converge to the desired point, the algorithm estimates the `FairShare` dividing `MaxAllocated` by the switch overload (sentence 23). Finally, note that the algorithm takes averages of the `Overload` and `MaxAllocated` to smooth the convergence to the equilibrium point (sentences 18∼22).

To have a pictorial view of the algorithm behavior, figures 10.7 and 10.8 show the queue length and source rates obtained with the network topology of figure 10.6. Sources are greedy with the weights shown in table 10.3. The averaging factor of the switch (`AVF` in the pseudo-code) is 0.6. The other switch and source parameters are the same as those of section 10.3.3. Sources start transmitting staggered 200 ms. Initially source S1 occupies alone the link L1. When source S2 starts transmitting, it receives triple bandwidth than S1, which produces the peak queue length in the switch X1. When the source S3 starts transmitting, it receives 4 times the bandwidth of source S2, which becomes bottlenecked at switch X2. The remaining bandwidth of source S2 in X1 is then given to S1. Finally, when source S4 becomes active, the switch X2 redistributes again the bandwidth according to the source weights.

We conclude that the algorithm quickly converges to the equilibrium, properly dividing the available bandwidth according to the source weights. Note that the equilibrium point is free of strong oscillations. However, during the transient phase, peaks in the queue length may grow, mostly if a source with a high weight becomes active. These peaks could be controlled by means of the RIF of the sources, or by introducing other switch controls, e.g. checking the queue length.

```
At each cell “cp” arrival: {
    EnqueueCell() ;
    if(cp.IsBackwardRMCell()) {
        cp.ER = min(cp.ER,
            min(cp.MCR + FairShare * VCTable[cp.VCI].Weight, cp.PCR)) ;
    } else {
        if(cp.IsForwardRMCell())
            MaxAllocated = max(MaxAllocated,
                (cp.CCR - cp.MCR) / VCTable[cp.VCI].Weight) ;
        count = count + 1 ;
        if(count >= MIC) EndOfInterval() ;
    }
}

EndOfInterval() {
    InputRate = count / (now() - StartOfInterval) ;
    Overload = max(InputRate - SumMCR, 1) / max(TCR - SumMCR, 1);
    if(Overload <= 1.  AND Overload >= 0.8)
        AvgOverload = AvgOverload + AVF * (Overload - AvgOverload) ;
    else AvgOverload = Overload ;
    if(MaxAllocated > 0.)
        AvgMaxAllocated = MaxAllocated + AVF * (MaxAllocated - AvgMaxAllocated) ;
    FairShare = AvgMaxAllocated / AvgOverload ;
    StartOfInterval = now() ; count = 0 ; MaxAllocated = 0.  ;
}
```

Figure 10.5: Weighted-switch algorithm pseudo-code.

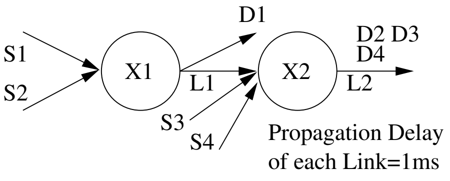


Figure 10.6: Network topology.

S1	S2	S3	S4
1	3	12	18

Table 10.3: Source weights.

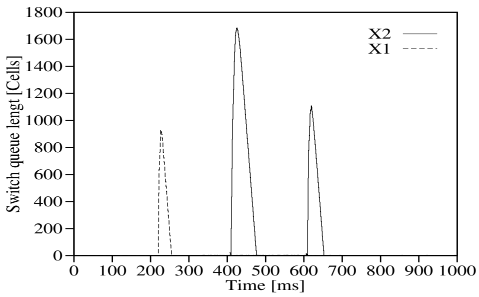


Figure 10.7: Switch queue lengths.

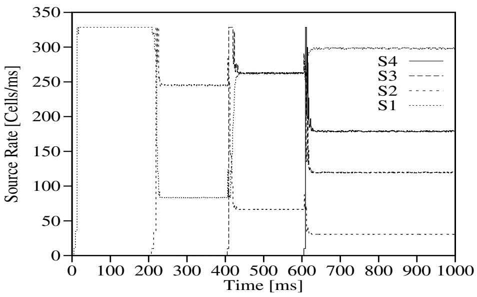


Figure 10.8: Source rates.

Chapter 11

ABR Support to TCP

11.1 Introduction

The Internet traffic supported by the TCP/IP set of protocols is currently the biggest part of non real time traffic. Therefore, ABR may be one of the Service Categories chosen to give support to the Internet traffic in ATM networks.

Since the ABR Service Category was specified, numerous studies have been done to analyze its inter-operation with the transport protocol used in Internet (TCP). Earlier studies of TCP over ATM can be found in [30] and [69]. These studies show that poor performance may be achieved due to the fragmentation of the TCP packets in the ATM network. The authors explore cell dropping policies to improve the performance. In [31] the authors show that the performance achieved with ABR may depend on the selection of the ABR parameters. In [44] the authors analyze the effect of the buffer size on the loss ratio and the performance achieved using ABR. In [55] a comparison is done between UBR and ABR for the interconnection of shared media LANs. In the scenarios analyzed in this study, the authors find that UBR outperforms ABR in terms of goodput. In [71] a proposal is done to improve the performance of ABR when the ABR control loop do no extend to the TCP end source. This proposal consist of doing a rate-to-window translation at the gateways.

In this chapter, simulations confronting different types of LAN interconnection are performed using ABR and the simpler UBR service category. The simulation results give guidelines about the TCP and ABR inter-operation and the benefits of using ABR over UBR.

The rest of the chapter is organized as follows. First, an overview of the TCP protocol is given in section 11.2. Section 11.3 describes the model of the devices that have simulated in this chapter. Then, the scenarios studied are explained and numerical results are given in section 11.4. Finally, section 11.5 gives some concluding remarks.

11.2 TCP Overview

TCP is a variable window protocol (see figure 11.1). The size of the window (`allowed_wnd` in the figure) limits the maximum number of segments, beyond the first unacknowledged one, that the TCP layer can send into the network. When this maximum is reached, the transmitter

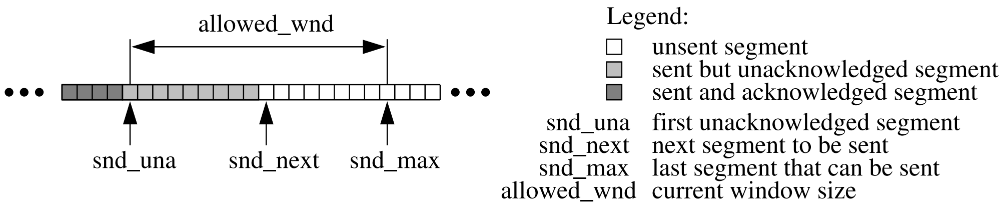

allowed_wnd
snd_una
snd_next
snd_max
Legend:
unsent segment
sent but unacknowledged segment
sent and acknowledged segment
snd_una first unacknowledged segment
snd_next next segment to be sent
snd_max last segment that can be sent
allowed_wnd current window size

Figure 11.1: TCP window mechanism.

stales and waits for new acknowledgments (acks) before sending new segments. A retransmission time-out is used to avoid a dead-lock waiting for acks.

Many of the current TCP implementations are based on four window adjustment algorithms developed by Van Jacobson [37, 38]:

- Slow start,
- congestion avoidance,
- fast retransmit and
- fast recovery.

These algorithms first appeared in the 4.3 BSD Reno release [74]. The TCP layer used in the simulations performed in this chapter implements these algorithms. In the following they are briefly described.

11.2.1 Slow Start and Congestion Avoidance

The size of the window used by the TCP module (`allowed_wnd` in figure 11.1) is upper-bounded by the *advertised window* which is a value set by the destination TCP module. Sending the full advertised window, however, may lead to buffer overflow along the transmission path. Lost segments have to be retransmitted, and this causes a throughput reduction. Having a small window may also reduce the throughput because of unnecessary waits for acks.

The Slow start algorithm is used to progressively increase the `allowed_wnd` of the TCP module seeking for a suitable size. This consists of the following rules [37, 74]:

- Add a congestion window (referred to as `cwnd`),
- the `allowed_wnd` is the minimum of the advertised window and the `cwnd`,
- when starting or restarting after a loss, `cwnd` is set to one segment,
- on each ack for new data, increase `cwnd` by one segment.

Note that with the former algorithm the `allowed_wnd` is opened one segment each time a segment is acknowledged, thus growing exponentially towards to the advertised window. If the transmission path is not able to store the full advertised window losses will occur. The congestion avoidance algorithm tries to stabilize the window to its optimum value. For stability reasons, the window is multiplicatively reduced when losses occur and additively increased otherwise. The time-outs are used as loss signals. Since time-outs are also used to restart the slow start algorithm, slow start and congestion avoidance algorithms are combined in the following way:

- A slow start threshold `ssthresh` is kept to switch between slow start and congestion avoidance algorithms,
- on a time-out, half of the current `allowed_wnd` is stored in `ssthresh` (this is the multiplicative decrease) and the slow start algorithm is initiated,
- slow start algorithm is applied until the `cwnd` reaches `ssthresh`. Then the `cwnd` is additively increased as `cwnd = 1 / cwnd` on reception of each ack.

Doing this way, the slow start opens the window quickly to what is assumed to be a safe operating point (half the window where losses were detected). Then, congestion avoidance is applied and the window is slowly increased probing for higher bandwidth to be available.

11.2.2 Fast Retransmit and Fast Recovery

In large bandwidth-delay product networks the previous algorithms may have a poor performance. This is because large windows may be required in such networks. Consequently, the TCP module may have many unacknowledged segments when a loss is detected by the time-out mechanism which may unnecessarily be retransmitted.

The fast retransmit algorithm exploits the fact that receiving consecutive duplicate acks is a likely indication that a segment has been lost. Therefore, the segment is transmitted without waiting for a time-out. After a fast retransmit, the following actions are applied (these are referred to as *fast recovery*):

- A congestion avoidance instead of a slow start algorithm is applied. This is because the full `allowed_wnd` may be in the transmission path when the fast retransmit is performed. Thus, it is not necessary to restart with the slow start algorithm.
- A new segment is allowed to be sent upon the reception of each duplicate ack. This is because the destination TCP module only generates an ack on the reception of a segment. Therefore, each ack indicates that a segment has left the transmission path and a new segment can be sent to take its place.

The fast retransmit and fast recovery algorithms are implemented together as follows [74, 75]:

- When the third consecutive duplicate ack is received: (i) the missing segment is retransmitted (this is the *fast retransmit*), (ii) `ssthresh` is set to max(`allowed_wnd`/2, 2), (iii) `cwnd` is set to `ssthresh` plus 3 segments. This inflates `cwnd` by the segments that have left the network.
- Each time a duplicate ack arrives, increment `cwnd` by one segment (for the corresponding segment leaving the network).

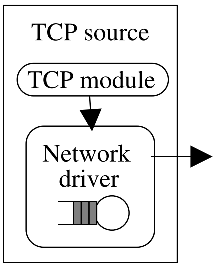


Figure 11.2: TCP source.

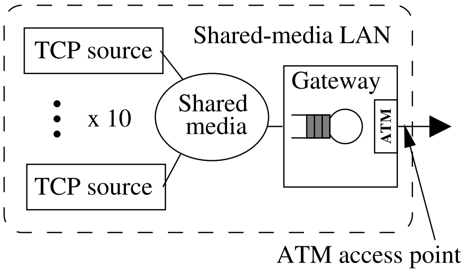


Figure 11.3: Shared-media LAN (SM-LAN).

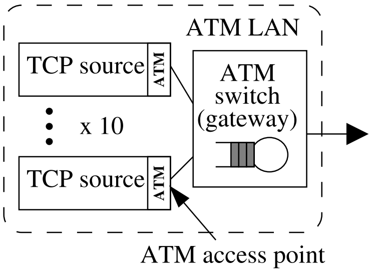


Figure 11.4: ATM-LAN.

- When the next ack that acknowledges new data is received, the `cwnd` is set to `ssthresh`. This ack should be the acknowledgment of the fast retransmit. Therefore, this step performs the congestion avoidance since the rate is reduced down to one-half the value it was when the packet got lost.

11.3 Simulation Model Description

This chapter considers the interconnection of Local Area Networks (LANs) with source and destination end systems running the TCP protocol. Two different cases are analyzed, each one with a different type of LAN. The two types of LANs are:

1. Shared media LANs (SM-LAN),
2. ATM-LANs.

In the following the assumptions made for the simulation of the TCP module and the LANs are described.

11.3.1 The TCP Module

The sources connected to the LANs run the Reno TCP protocol described in section 11.2 with the following parameters:

- Timer granularity of 100 ms.
- We assume that the destination always advertises a window size of $2^{16} - 1$ bytes (this is the maximum advertised window if the scale option is not used [74]).

We use the retransmission time-out mechanism described in [37]. Figure 11.2 shows the model used for the TCP sources. A greedy TCP module (which has always segments ready to send) is assumed. This TCP module passes the maximum number of segments allowed by the TCP window to the network driver. The segments are immediately given to the driver after the TCP module initialization, when an ack allows the transmission of more segments, or when a time-out expires.

For the network driver we have considered an infinite queue (i.e. with no losses) which stores the segments received by the TCP module until they are sent into the LAN. Furthermore, the TCP destination module sends an ack for each received segment. For sake of simplicity, we have not considered queuing delays but only the propagation delay in the return path of the acks. Additionally, a small random component in the return path delay of the acks has been added in order to avoid phase effects.

11.3.2 The LAN

Figures 11.3 and 11.4 show the model we have considered for the Shared media LAN (SM-LAN) and ATM-LAN respectively.

In the SM-LAN we simulate the transmission media with a 100 Mbps server that randomly polls the network driver of the TCP sources. All segments are assumed to be 1500 bytes long. The segments are stored into the queue of the gateway after leaving the shared media. The gateway is connected to the ATM network by a 155 Mbps link (all ATM links considered in the simulations are 155 Mbps). We assume that the ATM access point at the gateway has an AAL5 (thus, the last cell of each segment is marked with the *end of frame* bit). We have taken the AAL5 to use 32 cells for each TCP segment transmission. Furthermore, a single VC is used by the gateway to convey all the TCP traffic generated by the LAN.

For the ATM-LAN the topology shown in figure 11.4 has been used. Now, the TCP sources are connected to an ATM switch. All links are 155 Mbps and the propagation delay between the sources and this switch is 1 μs. In this case the ATM access point is located at the network driver of the TCP sources, thus, one VC per TCP source is used. As in the SM-LAN, we take an AAL5 using 32 cells per TCP segment.

Since we always use 32 cells per TCP segment at the ATM access point, and the advertised window is $2^{16}-1$ bytes (see section 11.3.1), we have taken a maximum window size of $\lfloor(2^{16}-1)/(32\cdot 48)\rfloor = 42$ segments. Furthermore, we shall use the following relation to compute the goodput:

$$\text{Goodput (bps)} = \frac{\text{Seq. num. increment of the transmitted segments} \cdot 32 \cdot 48 \cdot 8}{\text{Observation time (sec)}} \tag{11.1}$$

11.4 Numerical Results

11.4.1 Simulation Topology

Figure 11.5 shows the network topology simulated in this chapter. In this network we have taken two LANs (each having 10 TCP sources) and a guaranteed rate source as background traffic (indicated as VBR in the figure). The VBR traffic has full priority over the traffic offered by the LANs. This VBR source consists of an ON-OFF source with a deterministic ON and OFF periods of 0.5 seconds each. This source starts transmitting cells at time = 0 sec and the transmission rate in the ON period is half of the link rate.

These sources share a common output link of an ATM switch representing a public network.

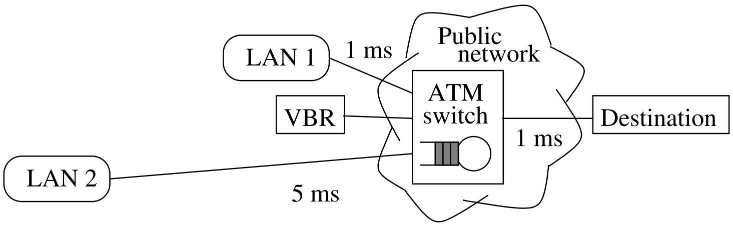


Figure 11.5: Simulation scenario.

The propagation delay between the LANs and the switch is 1 ms for LAN 1 and 5 ms for LAN 2. The delay between the switch and the destination is 1 ms.

11.4.2 Simulation Scenarios

Eight different simulations have been performed. These have been obtained combining the following items:

- Two different types of LANs: SM-LANs and ATM-LANs (see figures 11.3 and 11.4).
- Two different Service Capabilities: UBR and ABR. These are performed at the ATM access point (see figures 11.3 and 11.4). In case of UBR we have taken a transmission rate at the ATM access point equal to the Link Cell Rate (LCR). In case of ABR, the ABR Source End System behavior [4] is applied at the ATM access point. Thus, cell transmission rate is given by the Allowed Cell Rate (ACR).

  For the ABR case the ERICA switch algorithm is used (see section 5.2.3) with parameters: Target Cell Rate (TCR) = 0.9 · Link Cell Rate, Measuring interval = 100 cells.
- Two different buffer sizes for the gateway and the ATM switch. In the first case we have used a gateway with a buffer size of 30 segments (equivalent to 960 cells) and an ATM switch at the public network with a buffer size ten times higher (9600 cells). We have also considered the opposite case, i.e. gateways and ATM switch at the public network with 9600 cells respectively 960 cells. Note that in any of the cases neither the gateway nor the public switch can store the maximum window size of all the TCP sources (to store the maximum window size of the 10 TCP sources of one of the LANs a buffer size of 420 segments would be needed).

  A tail discarding policy at the ATM switches is assumed. Thus, when a cell is discarded because of buffer overflow, the following cells of the VC are discarded up to the cell with the end of frame bit set. In case of the SM-LAN the full segment is discarded if buffer overflow occurs at the gateway.

In each simulation all the TCP sources in the LANs start transmitting at time = 0 sec, and the statistics are taken averaging over the first minute of transmission time. Table 11.1 summarizes the results of the simulations. What we call gateway in table 11.1 represents the gateway of the SM-LAN case and the switch of the ATM-LAN respectively. For the TCP sources of each LAN the table gives: (i) The goodput, obtained by adding the goodput of the 10 TCP sources of the LAN computed with formula (11.1). (ii) The *Segment loss ratio* obtained as the number

			SM-LAN				ATM-LAN			
			UBR	ABR	UBR	ABR	UBR	ABR	UBR	ABR
		Switch buffer size (cells)	960		9600		960		9600	
		Gateway buffer size (cells)	9600		960		9600		960	
	Public network switch	Peak queue length (cells)	960	800	9600	1265	960	960	9600	1101
		Segment loss ratio (%)	6.13	0	1.55	0	5.02	0.05	1.81	0
		Load	0.85	0.90	0.99	0.90	0.93	0.90	0.99	0.90
LAN 1	Gateway	Peak queue length (cells)	32	9600	32	960	8999	20	960	20
		Segment loss ratio (%)	0	0.25	0	3.54	0	0	0.00	0
	TCP sources	Goodput (Mbps)	67.17	44.52	63.52	46.58	92.10	44.62	91.09	44.51
		Segment loss ratio (%)	6.20	0.25	1.33	3.54	4.72	0.02	0.89	0
		Segment retx ratio (%)	6.83	0.36	2.14	3.75	5.37	0.02	1.37	0
		Time-out ratio (%)	0.62	0.02	0.17	0.72	0.40	0	0.07	0
LAN 2	Gateway	Peak queue length (cells)	32	9600	32	960	547	20	547	20
		Segment loss ratio (%)	0	0.35	0	2.58	0	0	0	0
	TCP sources	Goodput (Mbps)	13.95	43.97	39.84	42.61	1.99	44.65	13.04	44.79
		Segment loss ratio (%)	5.84	0.35	1.90	2.58	16.87	0.07	7.75	0
		Segment retx ratio (%)	6.52	0.83	2.60	2.81	17.72	0.07	8.93	0
		Time-out ratio (%)	0.99	0.04	0.30	0.44	6.74	0.00	1.94	0
		Efficiency (%)	77.04	84.04	98.17	84.71	89.36	84.79	98.90	84.81
		Fairness index	0.63	0.99	0.94	0.99	0.52	1.00	0.63	0.99

Table 11.1: UBR and ABR TCP confront.

of segments lost by the 10 TCP sources of the LAN over the segments transmitted by these sources. (iii) The *Time-out ratio* and *Segment retx ratio* computed similarly to the *Segment loss ratio*, but using the number of Time-outs (respectively, retransmissions) over the transmitted segments.

Ideally, the *Segment retx ratio* should be equal to the *Segment loss ratio* (i.e. only the segments that are lost are retransmitted). Therefore, confronting both measures gives an idea of the segments that are unnecessarily retransmitted.

Table 11.1 gives the efficiency of the TCP sources as the overall goodput divided by the maximum achievable goodput. Since the VBR source occupies 25 % of the link bandwidth, the efficiency has been computed as:

$$\text{Efficiency} = \frac{\text{Overall goodput (Mbps)}}{155 \cdot 0.75 \cdot 48/53} \tag{11.2}$$

Table 11.1 gives also the fairness index computed applying the formula (4.5) to the goodput, i.e. computed as $\left(\sum_i x_i\right)^2 / \left(n \sum_i x_i^2\right)$, where $x_i = \tilde{x}_i / \hat{x}_i$, being $\tilde{x}_i$ the goodput achieved by source i and $\hat{x}_i$ the fair goodput for that source. Since all the TCP sources in our framework should get the same fair goodput, we have used the formula:

$$\text{Fairness index} = \frac{\left(\sum_i \text{goodput of source } i\right)^2}{n \sum_i \left(\text{goodput of source } i\right)^2} \tag{11.3}$$

11.4.3 Analysis

From the results shown in table 11.1 we derive the following conclusions:

SM-LAN: In UBR the segment losses are located at the public switch and in ABR at the gateway of the LANs. These is because in UBR the gateway always transmits at the LCR. Since the LCR is 155 Mbps and the transmission rate in the SM-LAN is 100 Mbps, the buffer at the gateway is always empty, storing at most 1 segment (32 cells). In ABR the transmission rate at the gateway is modulated by the available bandwidth at the public switch, therefore, the buffer is filled up when the gateway transmission rate is lower than 100 Mbps.

Consequently, using UBR the lower is the buffer size at the public switch the higher is the segment loss ratio. Using ABR the same reasoning is valid but with the buffer size at the gateway. This is confirmed by the results shown in table 11.1. Furthermore, the table shows that the overall efficiency obtained with UBR has a notable increment when increasing the buffer size at the public switch (because the loss ratio is reduced). However, when using ABR the overall efficiency remains nearly constant despite of the difference on the loss ratio.

The fairness index shows that while a good share is achieved with ABR, it is not like that with UBR. In fact the goodput results indicate that the TCP sources in LAN 1 get a higher bandwidth than the TCP sources in LAN 2 in all the cases. However, with UBR this difference is much larger, especially, when using a smaller buffer size at the public switch. In the next section this unfairness problem will be further investigated.

ATM-LAN: In this scenario the ATM access point is located at the TCP sources (as shown in figure 11.4). The consequence for the ABR sources is that a nearly zero loss ratio is achieved in all the cases. This is because the ABR control loop extends to the source avoiding that buffer overflow occurs at any switch. For the UBR sources the bottleneck located at the public switch is again responsible of most of the losses.

Looking at the efficiency results, table 11.1 shows similar figures as in the SM-LAN scenario. However, the unfairness produced using UBR is now increased. The table shows that using UBR the TCP sources of LAN 1 have much lower loss, retransmission and time-out ratios, and take most of the available bandwidth.

11.4.4 Dynamics of TCP over ABR and UBR

In this section traces of the simulations summarized in table 11.1 are shown in order to further investigate the interactions between the TCP protocol and the service category used in the ATM network (UBR and ABR). First, figures giving a general overview of the transmission process are analyzed. Then, more detailed traces of each scenario are presented.

General overview: Figures 11.6∼11.9 show the evolution of the sequence number of the segments sent during the first 10 seconds of the simulation. These figures depict the results obtained for the TCP sources of LAN 1 and LAN 2 in the cases of table 11.1 having a buffer size of 960 cells and 9600 cells at the public switch and the gateway respectively (this is indicated as "`SX=960, GW=9600`" in the title of the figures).

Figures 11.6∼11.8 show that the number of transmitted segments varies among the TCP sources. This is due to the random distribution of the segment losses. Figure 11.9, instead, shows that the segment transmission is exactly the same for all TCP sources. This is because when using ABR in the ATM-LAN scenario there are no segment losses and the switches equally divide the available bandwidth among the sources (see table 11.1).

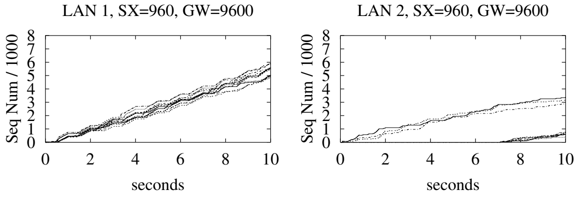


Figure 11.6: SM-LAN, UBR.

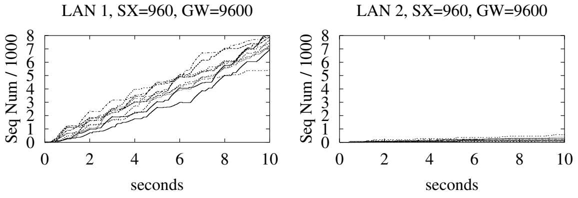


Figure 11.7: ATM-LAN, UBR.

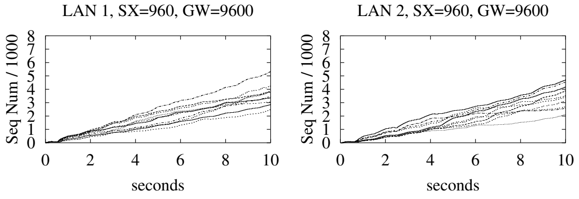


Figure 11.8: SM-LAN, ABR.

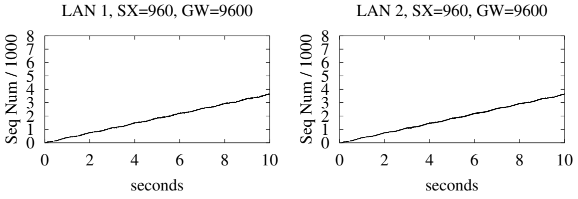


Figure 11.9: ATM-LAN, ABR.

Furthermore, figures 11.6 and 11.7 show that the TCP sources of LAN 1 achieve a much higher segment transmission rate than the TCP sources of LAN 2. This confirms the unfairness problem mentioned in the previous section. Finally, figure 11.6 shows that an unfairness problem arises among the TCP sources of the same LAN. In fact, some of the TCP sources of LAN 2 achieve a much higher segment transmission than the others.

Scenario with SM-LANs and UBR: In order to see the interaction between TCP and UBR in the SM-LAN scenario, figures 11.10∼11.15 show the following traces:

- Figure 11.10 shows a zoom of 80 ms of the trace of one of the TCP sources of LAN 1 depicted in figure 11.6. The figure shows the instants when the TCP module passes the segments to the TCP driver (as would be recorded by the `tcpdump` utility in a real scenario [74]), the instants when a segment loss occurs, and the instants when acks arrive at the TCP module.
- For the same TCP source as figure 11.10, figure 11.11 shows the segment transmission and losses during the first 5 seconds of simulation (the sequence number of this figure is plotted module 1000 for sake of clarity). Furthermore, the `cwnd` and `ssthresh` are plotted in the lower side of the figure. Remember that `ssthresh` is the threshold at which the slow start algorithm changes to congestion avoidance.
- Figure 11.12 shows the transmission rate measured averaging over periods of 10 ms at the output of the gateway of LAN 1 (in the upper side of the figure) and LAN 2 (in the lower side of the figure). For sake of comparison, the background traffic offered by the VBR source is superimposed in dots.
- For the same TCP source as figures 11.10 and 11.11, figure 11.13 shows the queue length built up at the buffer of the network driver of the source (in the upper side of the figure). In the lower side of the figure the same trace is shown for one of the sources of LAN 2.
- Figure 11.14 shows the queue length at buffer of the gateway of LAN 1 (in the upper side of the figure) and LAN 2 (in the lower side of the figure).

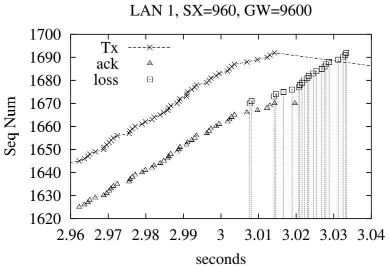


Figure 11.10: Trace of 80 ms. SM-LAN, UBR.

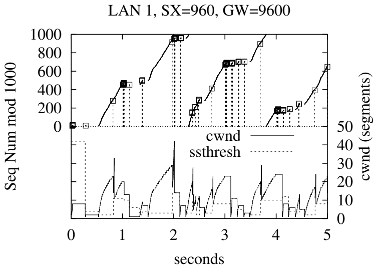


Figure 11.11: Trace of 5 sec. SM-LAN, UBR.

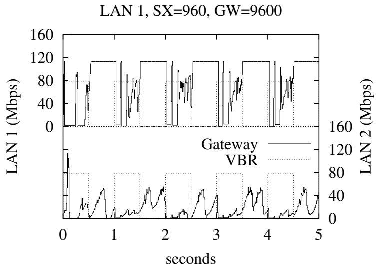


Figure 11.12: Transmission rate. SM-LAN, UBR.

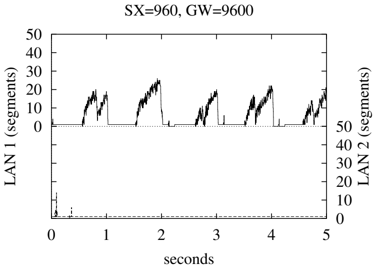


Figure 11.13: Driver buffer occupancy. SM-LAN, UBR.

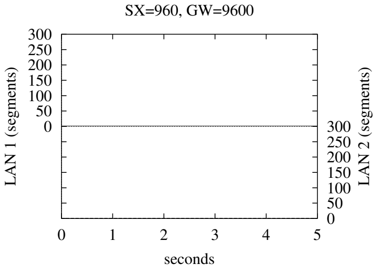


Figure 11.14: Gateway buffer occupancy. SM-LAN, UBR.

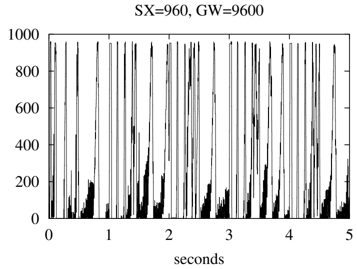


Figure 11.15: Switch buffer occupancy. SM-LAN, UBR.

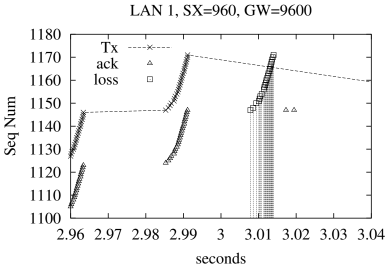


Figure 11.16: Trace of 80 ms. ATM-LAN, UBR.

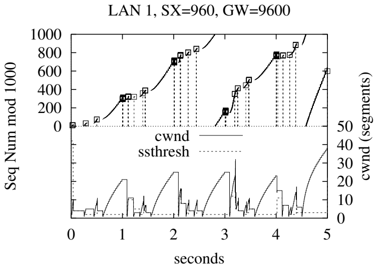


Figure 11.17: Trace of 5 sec. ATM-LAN, UBR.

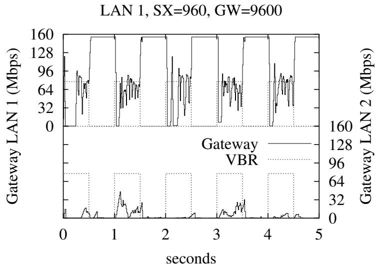


Figure 11.18: Transmission rate. ATM-LAN, UBR.

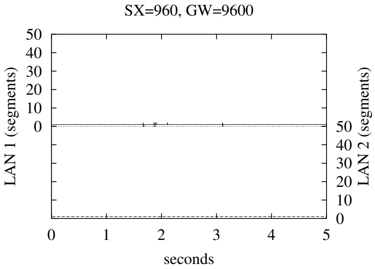


Figure 11.19: Driver buffer occupancy. ATM-LAN, UBR.

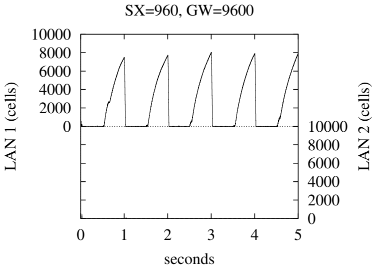


Figure 11.20: Gateway buffer occupancy. ATM-LAN, UBR.

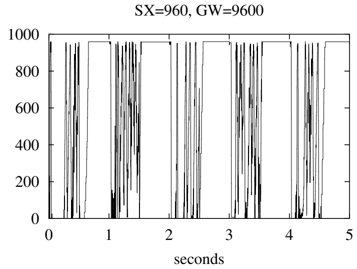


Figure 11.21: Switch buffer occupancy. ATM-LAN, UBR.

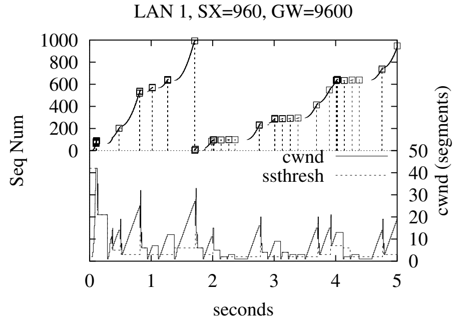


Figure 11.22: Trace of 5 sec. SM-LAN, UBR, LAN 2.

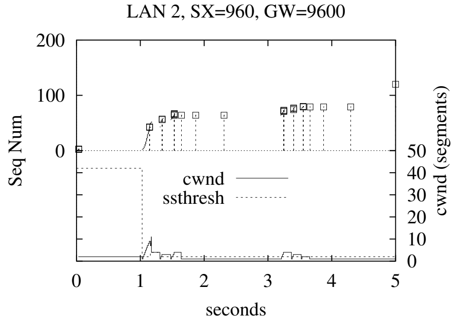


Figure 11.23: Trace of 5 sec. ATM-LAN, UBR, LAN 2.

- Finally, figure 11.15 shows the queue length at buffer of the public switch.

The same traces are also shown for the ATM-LAN scenario using UBR (figures 11.16∼11.21), for the SM-LAN scenario using ABR (figures 11.24∼11.29), and for the ATM-LAN scenario using ABR (figures 11.30∼11.35). In the following these traces are explained and confronted.

Figure 11.10 shows a delay of ≈20 ms since the TCP module transmits a segment until it receives the corresponding ack. This delay is mainly due to the propagation delay and the waiting time in the queues. In this case the queue at the gateway stores at most one segment (see figure 11.14) and the public switch is nearly empty before time = 3 sec. However, figure 11.13 shows that the queue at the network driver stores ≈15 segments which approximately need $(15 \cdot 1500 \cdot 8)/(100 \cdot 10^3/10) = 18$ ms to be served by the 100 Mbps shared media LAN. Therefore, we conclude that the main part of the round trip delay is located at the queue of the network driver of the TCP source.

Surprisingly, figure 11.13 shows that the buffer at the network driver is filled during the OFF periods of the VBR source and it is emptied during the ON periods (note from figure 11.12 that the VBR source is ON the first half of each second). This is because during the OFF periods, the TCP sources of LAN 1 increase the `cwnd` rapidly (as shown if figure 11.11), and thus, the transmission rate, filling the buffer of the driver. Figure 11.12 shows that the transmission rate is much higher at the gateway of LAN 1 than at the gateway of LAN 2. This confirms the unfairness problem mentioned in section 11.4.3.

This unfairness effect is due to different propagation delays in the transmission path (see figure 11.5). Since sources of LAN 1 have a shorter propagation delay than sources in LAN 2, sources of LAN 1 increase the transmission rate at a higher speed than the sources of LAN 2. This can be seen confronting the trace of `cwnd` in figures 11.11 and 11.22. Note that figure 11.22 shows the same traces as figure 11.11, but for one of the sources of LAN 2. Figure 11.11 shows that each time the `cwnd` of the LAN 1 source grows, it does so initially very fast, and progressively slows down due to the increase of the round trip time as a consequence of the queue built up at the network driver (see figure 11.13). Figure 11.22 shows that the `cwnd` of the LAN 2 source starts growing more slowly. This allows the sources of LAN 1 to recover faster than sources of LAN 2 after a loss, and thus, achieving a higher transmission rate.

Scenario with ATM-LANs and UBR: Figures 11.16∼11.21 show similar results as 11.10∼11.15. A difference arises in the buffer occupancy of the network driver of the TCP source (figures 11.13 and 11.19) and the gateway of LAN 1 (figures 11.14 and 11.20). Note that the queue that was built up at the network driver in the SM-LAN scenario (figure 11.13), is built up at the gateway in the ATM-LAN scenario (figure 11.20). This is because the bottleneck due to the shared media in the SM-LAN is shifted to the output link of the gateway in the ATM-LAN scenario.

Figure 11.18 shows that the gateway of LAN 1 achieves a much higher transmission rate than the gateway of LAN 2. This unfairness is motivated by the same reason as in the SM-LAN scenario, i.e. because sources of LAN 1 are favored by having a shorter propagation delay than sources of LAN 2. However, figures 11.12 and 11.18 show that the unfairness problem observed in the SM-LAN scenario is increased in the ATM-LAN scenario. This is because in the SM-LAN scenario the sources of the LANs are limited by the 100 Mbps of the shared media. In the ATM-LAN scenario this constrain does not exist and sources of LAN 1 can take the full 155 Mbps of the link of the public switch shared with the sources of LAN 2.

The unfairness problem explained in the former paragraph is also illustrated in figure 11.23. This figure depicts the same traces as figure 11.17, but for one of the sources of LAN 2. The figure shows that consecutive retransmissions of the same segment are lost. Note that the backoff algorithm of TCP duplicates the value of the timer each time a time-out is triggered. With these figures we conclude that, in the ATM-LAN scenario using UBR, the sources of LAN 2 are nearly blocked by sources of LAN 1.

Scenario with SM-LANs and ABR: Figures 11.24∼11.29 are analogous to 11.10∼11.15 but for the SM-LAN over ABR scenario. Figure 11.26 shows the effect of the ABR control loop between the gateways and the switch at the public network. Comparing with figure 11.12 we can see that each gateway receive the same bandwidth in ABR, while in UBR the gateway of LAN 1 is much more favored. Furthermore, figure 11.12 shows higher oscillations, demonstrating the finer control achieved with ABR. Figure 11.28 shows that the queue is built up at the edge of the ABR control loop, i.e. the gateway in this scenario.

Scenario with ATM-LANs and ABR: Finally, figures 11.30∼11.35 show the ATM-LAN over ABR case. Now the ABR control loop extends up to the network driver of the TCP sources. This motivates that the queue length is built up at the network driver of the sources, as shown in figure 11.33.

Figure 11.30 shows a delay of ≈80 ms and then ≈160 ms between the transmission of a segment by the TCP module and the reception of the corresponding ack. Note that this delay is higher than in any of the previous scenarios. The explanation for this long round trip time is that nearly the full window used by the TCP module is stored at the network driver of the TCP source, as shown in figure 11.33. Since the switch at the public network divides the available bandwidth among the TCP sources, each receives $0.9 \cdot 155/20 \approx 7$ Mbps when the VBR source is silent. Therefore, a segment arriving at the network driver is likely to wait $(42{\cdot}32{\cdot}53{\cdot}8)/(7{\cdot}10^3) \approx 80$ ms before leaving the driver when the VBR source is silent and 160 ms when is active, as shown in figure 11.30.

Note that in this scenario the ABR control mechanism completely substitutes the flow control performed by TCP. Indeed, figure 11.31 shows how TCP initially tries to probe for higher bandwidth continuously increasing the `cwnd` up to the advertised window. Then, the `cwnd` remains constant because no losses are produced. This could be a problem since, as mentioned

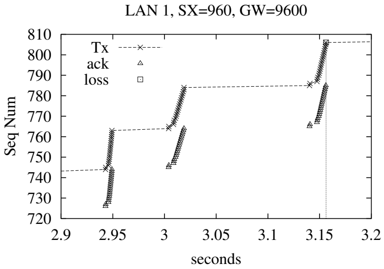


Figure 11.24: Trace of 300 ms. SM-LAN, ABR.

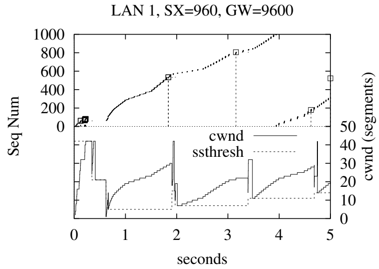


Figure 11.25: Trace of 5 sec. SM-LAN, ABR.

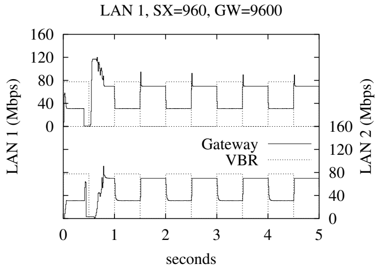


Figure 11.26: Transmission rate. SM-LAN, ABR.

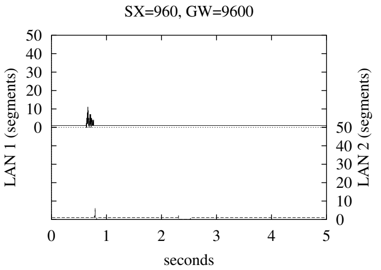


Figure 11.27: Driver buffer occupancy. SM-LAN, ABR.

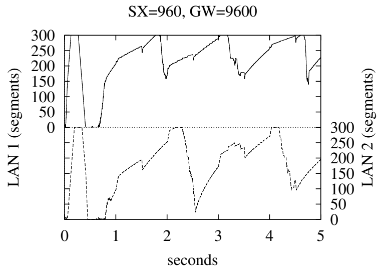


Figure 11.28: Gateway buffer occupancy. SM-LAN, ABR.

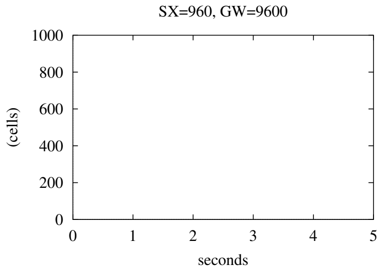


Figure 11.29: Switch buffer occupancy. SM-LAN, ABR.

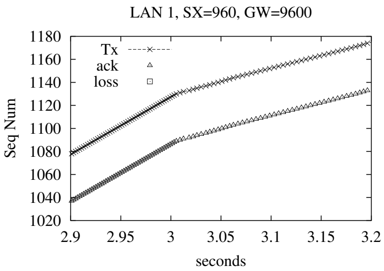


Figure 11.30: Trace of 300 ms. ATM-LAN, ABR.

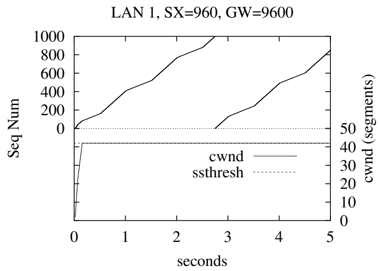


Figure 11.31: Trace of 5 sec. ATM-LAN, ABR.

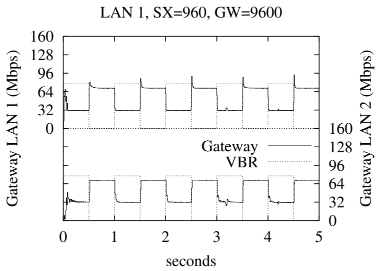


Figure 11.32: Transmission rate. ATM-LAN, ABR.

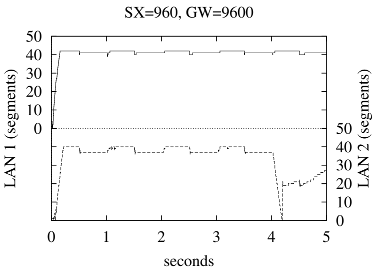


Figure 11.33: Driver buffer occupancy. ATM-LAN, ABR.

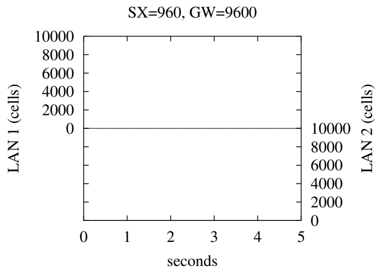


Figure 11.34: Gateway buffer occupancy. ATM-LAN, ABR.

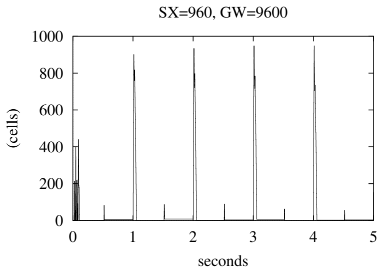


Figure 11.35: Switch buffer occupancy. ATM-LAN, ABR.

in the former paragraph, the ABR control loop makes the full window used by the TCP module to be stored at the network driver of the source. Therefore, if the network driver at the TCP source is not able to store the full advertised window, segments would be lost before leaving the TCP source!.

11.5 Conclusions

This chapter investigates by simulation the interactions between the TCP protocol used in the Internet and the ABR and UBR Service Categories standardized for ATM networks. These interactions may depend on many parameters related to the TCP protocol, the ATM Service Categories and the network. Therefore, the study has been restricted to a descriptive analysis, giving results that show the factors that may mostly influence the traffic behavior. To do so, averages and plots of the traces of the most representative parameters are shown in the chapter. These include goodput, buffer occupancy, allowed window of the TCP module etc.

We have considered the interconnection of Local Area Networks (LANs). Two cases are analyzed: one with Shared media LANs (SM-LAN), and another with ATM-LANs. With SM-LANs the access to the ATM network is located at the gateway of the LAN, and all the TCP traffic is conveyed over the same VC. With ATM-LANs, instead, the access to the ATM network is located at the network driver of the TCP sources, and each one has its own VC.

Our study allows to give some general guidelines about the interactions between TCP and the service category used in the ATM network (UBR or ABR). These are described in the following.

- Using UBR the queue length built up at the buffers tends to be pushed forward along the transmission path. This is because the maximum transmission rate is used to send the cells using this service category. Therefore, buffers located at the bottleneck points are filled up to the maximum capacity and losses occur.

- Using ABR the queue tends to be built up at the access point to the ATM network. This is due to the rate control performed by the ABR control loop. Consequently, the buffer at the gateway is filled up with SM-LANs, and the buffer at the network driver of the sources is filled using ATM-LANs.

- Higher segment loss ratios are obtained using UBR. This is because UBR exclusively depends on the flow control performed at the TCP level in order to adapt the transmission rate of the sources to the available bandwidth. This rate control has a longer feedback delay than the rate control performed at the ATM level with ABR.

- Despite of higher losses, higher overall goodput is generally achieved using UBR than ABR. This is because ABR keeps a percentage of the link load at the switches free, in order to be able to drain the queues after overload periods. Furthermore, a part of the available capacity is consumed by the transmission of the RM-cells. We conclude that these overheads introduced in ABR have a higher impact on the goodput than the higher segment losses which occur in UBR.

- Fairness problems easily arise in UBR. This problem is eliminated using ABR.

- ABR tends to eliminate the flow control mechanism of TCP. This is because with TCP this control is based on segment losses, which ABR tries to avoid. This is evident in the

ATM-LAN scenario where the ABR control loop extends to the TCP sources. In this case the TCP increments the allowed window to its maximum value and remains constant, since no losses occur. The consequence is that the full window is stored at the buffer of network driver of the source. This may be a problem if the buffer at the source is not able to store this maximum window size, since losses would occur at the source.

Summary and Outlook

Data traffic has emerged as a big challenge for the standardization of traffic management mechanisms in ATM networks. This fact can be derived by the diversity of SCs/ATCs defined for this kind of non real time traffic by the ATM Forum and the ITU-T respectively. UBR, ABR, GFR, ABT are ready standards and new SCs/ATCs are likely to be introduced, e.g. the Controlled Transfer (CT) ATC [36].

Many reasons have influenced this high number of SC/ATCs. First, the loose requirements that can be found in this kind of traffic. Generally speaking, data traffic can be defined by having an absence of tight time delay restrictions. This property can also be identified as the ability of this traffic to adapt its transmission rate to the available network resources. Some authors call the traffic having this characteristic *elastic* traffic [72, 45]. Internet traffic is an example of elastic traffic, however, differences may be found depending on the application. For example, a web page which is accessed for immediate attention tolerates delays of several seconds, while an application as e-mail can accept delays of minutes or hours.

The loose requirements of data traffic motivated the standardization of the firsts SCs/ATCs (UBR, ABR and ABT) focusing on different goals. ABT keeps common functionalities with the DBR ATC. Furthermore the source operations are simple. ABT principles have been explained in chapter 3. Remember that this ATC allows the renegotiation of the transmission rate, behaving similar to DBR when a new transmission rate is accepted. These characteristics could allow an easy deployment of ABT equipment.

The ATM Forum pushed towards two SCs based on different principles: UBR and ABR. UBR is the simplest SC and the cheapest to implement. UBR do not offer any QoS and may suffer from efficiency and severe fairness problems as shown in chapter 11. ABR was defined with ambitious objectives: high network efficiency, fairness and inter-operability of different ABR switch mechanisms. Since the ATM Forum specification of ABR appeared in April 1996, an intensive research has been carried out about this promising SC.

The major part of this PhD has been devoted to ABR. Instead of focusing on one aspect of ABR, the main research topics involved in ABR have been covered, namely: (i) switching mechanisms, (ii) conformance definition, (iii) charging, (iv) ABR support to TCP traffic. In the following the main conclusions are summarized.

Maybe, switch algorithms have been the most investigated topic of ABR. This has happened because the specification of ABR given by the ATM Forum allows a diversity of switch algorithms to be implemented. These range from the simplest binary switches [80, 57] to the more complex ER switches [42, 32, 46, 11, 68, 43, 77, 41]. In chapter 5 three of these switch algorithms are analyzed by means of simulation, showing the different degree of performance and complexity that can be achieved. The behavior of ER switches is also addressed in chapter 6. However, in this chapter real traces obtained with a commercial ER switch are shown. Chapter 7 completes the study of ABR switches discussing the difficulties that ER algorithms afford in order to fairly divide the available bandwidth among the contending sources. The chapter proposes to further exploit the functionalities of the CCR field of RM-cells to simplify the switch algorithms.

Chapters 8-9 analyze the conformance definition that has been standardized for ABR: the Dynamic Generic Cell Algorithm (DGCRA). The conformance definition is the formalism established to decide whether the source transmits according to the traffic contract. The conformance definition may be implemented in the Usage Parameter Control (UPC) for policing. Chapter 8 gives a detailed description of the DGCRA. Furthermore, traces obtained by simulation are depicted showing that the algorithm given by the ATM Forum has a decreasing accuracy of the rate conformance with increasing feedback delay. A *UPC based on the CCR* is proposed to solve this drawback.

Chapter 9 studies by means of two analytical approaches the parameter dimensioning of the DGCRA. Numerical results calculated with the analytical models are also obtained by simulation for validation. The analytical approaches are based on a novel queuing model of the DGCRA presented in this chapter. The first analytical approach is based on a renewal assumption of the cell inter-arrival process at the UPC. This approach gives a simple but coarse approximation of the cell rejection probability at the UPC. The second analytical method consists of a Markov chain which accurately describes the stochastic variables involved in the queuing model of the DGCRA. The Markov chain is solved applying the matrix geometric technique. The complexity of this mathematical approach only allows to investigate a simple network topology. However, the accuracy of the model allows to take into account the influence of the delay bounds τ_2 and τ_3 that are negotiated with the DGCRA. This study shows that a major degradation of the cell rejection probability may be obtained if these delay bounds are not properly set.

Chapter 10 investigates some ABR charging schemes. Charging may have a decisive impact on the deployment, success and growth of a network. In fact, the research community has paid a great attention to this topic in recent years. Furthermore, pricing may be an essential condition for the users when submitting traffic. Some authors have used this fact to propose congestion control mechanisms based on a dynamic pricing. In such schemes, prices vary according to the demand of network resources by the sources. New prices are conveyed to the sources by means of a feedback mechanism. This charging scheme seems to fit well with ABR, since the RM-cells can be used to dynamically communicate the prices. In chapter 10 a dynamic pricing scheme is proposed and an analytical model is used to find out the evolution of the prices. Additionally, several charging schemes are confronted by simulation. This comparison shows that the dynamic pricing gives the best expected charging.

Finally, chapter 11 investigates by simulation the support of ABR to the traffic generated with the TCP protocol used in the Internet. Currently, the data communications are dominated by the Internet traffic transported by a variety of networks. The deployment of ATM technology has been located in the backbone networks and the end-to-end ATM systems appear remote. In fact, it is not clear whether the universal multi-service network will be built on the Internet rather than the B-ISDN. An example of the increasing attention given to Internet has been the recent standardization of the Guaranteed Frame Rate (GFR) SC by the ATM Forum. This SC is specifically addressed to transport the kind of traffic generated by the Internet. However, the study of GFR goes beyond the scope of this PhD. Simulations performed in chapter 11 confront the transport of TCP traffic in different scenarios using ABR and the simpler UBR SC. The main conclusion is that ABR can solve the severe fairness problems that can arise using UBR.

Acronyms

AAL	ATM Adaptation Layer
ABR	Available Bit Rate
ABT	ATM Block Transfer
ABT/DT	ATM Block Transfer / Delayed Transfer
ABT/IT	ATM Block Transfer / Immediate Transfer
ACR	Allowed Cell Rate
ADTF	ACR Decrease Time Factor
ATC	ATM Transfer Capabilities
ATM	Asynchronous Transfer Mode
AUU	ATM layer User to ATM layer User
BT	Burst Tolerance
B-ISDN	Broadband ISDN
CAC	Connection Admission Control
CBR	Constant Bit Rate
CCR	Current Cell Rate
CDV	Cell delay variation
CDVT	Cell Delay Variation Tolerance
CLR	Cell Loss Ratio
CLP	Cell Loss Priority
CPN	Customer Premises Network
CRC	Cyclic Redundancy Check
CTD	Cell Transfer Delay
DBR	Deterministic Bit Rate
DES	Destination End System
DGCRA	Dynamic Generic Cell Rate Algorithm
EFCI	Explicit Forward Congestion Indicator
EPRCA	Enhanced Proportional Rate Control Algorithm
ER	Explicit Rate
ERICA	Explicit Rate Indication for Congestion Avoidance
FIFO	First in first out
FRP	Fast Reservation Protocol
GCRA	Generic Cell Rate Algorithm
GFC	Generic Flow Control
GFR	Guaranteed Frame Rate
HEC	Header Error Check
IBT	Intrinsic Burst Tolerance
ICR	Initial Cell Rate
IP	Internet Protocol
ISDN	Integrated Services Digital Network
ISO	International Organisation for Standardisation
ITU	International Telecommunications Union
ITU-T	ITU Telecommucation sector
LAN	Local Area Network
LVST	Last Virtual Scheduled Time
maxCTD	Maximum Cell Transfer Delay
MBS	Maximum Burst Size
MFS	Maximum Frame Size
MCR	Minimum Cell Rate
NI	No Increase bit
nrt-VBR	Non-Real-Time Variable Bit Rate
OAM	Operation And Maintenance
PCR	Peak Cell Rate

PTI	Payload Type Indicator
QoS	Quality of Service
RDF	Rate Decrease Factor
RIF	Rate Increase Factor
RM	Resource Management
RM-cell	Resource Management Cell
RTT	Round Trip Time
rt-VBR	Variable Bit Rate with real-time constraints
SBR	Statistical Bit Rate
SC	Service Category
SCR	Sustainable Cell Rate
SES	Source End System
TCP	Transmission Control Protocol
TCR	Target Cell Rate
UBR	Unspecified Bit Rate
UDP	User Datagram Protocol
UNI	User Network Interface
UPC	Usage Parameter Control
VBR	Variable Bit Rate
VC	Virtual Connection
VCI	Virtual Channel Identifier
VP	Virtual Path
VPI	Virtual Path Identifier

Index

Bibliography

[1] *AC-1000, ATM Protocol Analyzer.*
`http://www.ablecommus.com, http://www.ablecom.co.jp`

[2] A. Arulambalam, X. Chen, and N. Anasri. "Allocating Fair Rates for Avilable Bit Rate Service in ATM Networks". *IEEE Communications Magazine*, pages 92–100, November 1996.

[3] A. Arulambalam, X. Chen, and N. Anasri. "An Intelligent Explicit Rate Control Algorithm for ABR Service in ATM Networks". In *Proc. of the ICC'97*, Montreal, Canada, 1997.

[4] ATM Forum Technical Committee Traffic Management Working Group. *"ATM Forum Traffic Management Specification Version 4.0"*, April 1996.

[5] ATM Forum Technical Committee Traffic Management Working Group. *"ATM Forum Traffic Management Specification Version 4.1"*, March 1999.

[6] *8000 Series ATM Switches.*
`http://www.atmsys.com`

[7] G.M. Bernstein and D.H. Nguyen. "Blocking Reduction in Fast Reservation Protocols". In *Proc. of the IEEE INFOCOM'94*, 9c.2, pages 1208–1215, 1994.

[8] C. Blondia and O. Casals. "Analysis of Explicit Rate Congestion Control in ATM Network-s". In *Proc. of the Australian Telecommunications Networks and Applications Conference (ATNAC'96)*, Melbourne, Australia, December 1996.

[9] P.E. Boyer and D.P. Tranchier. "A reservation principle with applications to the ATM traffic control". *Computer Networks and ISDN Systems*, 24, 321–324, 1992.

[10] O. Casals, C. Blondia, L. Cerdà, and B. Van Houdt. "Congestion Control and Charging for the ABR Service Category in ATM Networks". In *Proc. of the SPIE Conf. on Performance and Control of Network Systems II*, volume 3530, pages 218–229, Boston, USA, November 1998.

[11] D. Cavendish, M. Gerla, and S. Mascolo. "ATM Rate Based Congestion Control Using a Smith Predictor: Implementation Issues". In *Proc. of the IFIP WG–6.2 First Workshop on ATM Traffic Management WATM'95*, pages 289–296, Paris, France, December 1995.

[12] L. Cerdà and O. Casals. "A Simulation Study of Switching Mechanisms for ABR Service in ATM Networks". Technical report, UPC-DAC-1996-21, September 1996.

[13] L. Cerdà and O. Casals. "Improvements and Performance Study of the Conformance Definition for the ABR Service in ATM Networks". In *Proc. of the ITC Specialists Seminar on Control in Communications*, pages 323–334, Lund, Sweden, September 1996.

[14] L. Cerdà and O. Casals. "Charging of the ABR Service in ATM Networks". In *Proc. of the 6th Open Workshop on High Speed Networks*, pages 111–118, Stuttgart, Germany, October 1997. University of Stuttgart.

[15] L. Cerdà and O. Casals. "Effective Usage of the CCR in the ABR Service Management". In *Proc. of the International Conference for Computer Communications (ICCC)*, pages 199–206, Cannes, France, November 1997.

[16] L. Cerdà and O. Casals. "CDVT Dimensioning in the ABR Service of ATM Networks". Technical Report UPC-DAC-1998-29, Politechnic University of Catalonia, September 1998.

[17] L. Cerdà and O. Casals. "Dynamic and Static Charging of the ABR Service in ATM Networks". In *Proc. of the GLOBECOM'98*, Sydney, Australia, November 1998.

[18] L. Cerdà and O. Casals. "Experimental Analysis of an ER Switch for the ABR Service in ATM Networks". In *Proc. of the IV Jornadas de Informática*, pages 449–458, Gran Canaria, Spain, July 1998.

[19] L. Cerdà and O. Casals. "Dimensioning of the CDVT in the Conformance Definition for the ABR Service in ATM Networks". In *Proc. of the* 5^{th} *Open European Summer School, EUNICE'99*, pages 71–76, Barcelona, Spain, September 1999.

[20] L. Cerdà and O. Casals. "Study of the TCP behavior over the ABR and UBR Services Categories in ATM Networks". Technical Report UPC-DAC-1999-37, Politechnic University of Catalonia, September 1999.

[21] L. Cerdà and O. Casals. "Charging of the ABR Service in ATM Networks: A numerical Example". *looking .forward, a Supplement to Computer*, Spring 1998.

[22] L. Cerdà, J. García, and O. Casals. "A Study of the Fairness of the Fast Reservation Protocol". In *Proc. of the IFIP TC6 3th Workshop on Performance Modelling and Evaluation of ATM Networks*, Bradford, UK, July 1995.

[23] L. Cerdà, B. Van Houdt, O. Casals, and C. Blondia. "Performance Evaluation of the Conformance Definition for the ABR Service in ATM Networks". *to be presented in Broadband Communications'99*, November 1999.

[24] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang. "Pricing in Computer Networks: Motivation, Formulation and Example". *IEEE/ACM Transactions on Networking*, 1(6), December 1993.

[25] C. Courcoubetis, V.A. Siris, and G.D. Stamoulis. "Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks". In *Proc. of the GLOBECOM'96*, London, UK., November 1996.

[26] J. Enssle, U. Briem, and H. Kröner. "Performance Analysis of Fast Reservation Protocols for ATM". In *Proc. of the IFIP TC6 2nd Workshop on Performance Modelling and Evaluation of ATM Networks*, pages 24/1–24/14, Bradford, UK., 1994.

[27] *European ACTS Project AC094 EXPERT.*
`http://www.elec.qmw.ac.uk/expert`

[28] EXPERT WP 4.1 and 4.2. *"Specification of Integrated Traffic Control Architecture"*, September 1996. AC094/EXPERT/WP(G)4/006.

[29] EXPERT WP 4.1 and 4.2. *"First Results from Trials of Optimized Traffic Control Features"*, March 1997. AC094/EXPERT/WP(G)4/010.

[30] C. Fang, H. Chen, and J. Hutchins. "A Simulation Study of TCP Performance in ATM Networks". In *Proc. of the GLOBECOM'94*, pages 1217–1223, 1994.

[31] C. Fang and A. Lin. "A Simulation Study of ABR Robustness with Binary-Mode Switches: Part II". ATM Forum contribution number 95-1328, October 1995.

[32] M. Gerla and S. Mascolo. "ATM Rate Based Congestion Control Using a Smith Predictor: an EPRCA Implementation Issues". In *Proc. of the IEEE INFOCOM'96*, 5b.2, pages 569–576, 1996.

[33] A. Gravey, J. Boyer, K. Sevilla, and J. Mignault. "Resource allocation for worst case traffic in ATM Networks". *Performance Evaluation*, 30, 19–43, 1997.

[34] R. Guerin. "UBR+ Service Category Definition". ATM Forum contribution number 96-1598, December 1996.

[35] ITU-T. *"Recommendation I.371, Traffic Control and Congestion Control in B-ISDN"*, July 1995.

[36] ITU-T SG/13 June'98 plenary meeting. *"Annex 1 - Q7 (ATM level traffic control)"*, June 1998.

[37] V. Jacobson. "Congestion Avoidance and Control". *ACM Computer Communication Review*, 18(4), 314–329, August 1988.
`ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z`

[38] V. Jacobson. "Modified TCP Congestion Avoidance Algorithm". end2end-interest mailing list, April 1990.
`ftp://ftp.isi.edu/end2end/end2end-interest-1990.mail`

[39] R. Jain. "Congestion Control and Traffic Management in ATM Networks: Recent Advances and a Survey". *Computer Networks and ISDN Systems*, November 1996.

[40] R. Jain, S. Kalyanaraman, R. Goyal, S. Fahmy, and R. Viswanathan. "ERICA Switch Algorithm: A Complete Description". ATM Forum contribution number 96-1172, August 1996.

[41] R. Jain, S. Kalyanaraman, R. Viswanathan, and R. Goyal. "A Sample Switch Algorithm". ATM Forum contribution number 95-0178R1, February 1995.

[42] P. Johansson and J.M. Karlsson. "Characteristics of an Explicit Rate ABR Algorithm". In *Proc. of the ITC Specialists Seminar on Control in Communications*, pages 309–321, Lund, Sweden, September 1996.

[43] L. Kalampoukas and A. Varma. "An efficient rate allocation algorithm for ATM networks providing max-min fairness". In *Proc. of the IFIP 6th International Conference on High Performance Networking*, pages 143–154, 1995.

[44] S. Kalyanaraman, R. Jain, S. Fahmy, R. Goyal, L. Fang, and S. Srinidhi. "Performance of TCP/IP over ABR Service on ATM Networks". In *Proc. of the GLOBECOM'96*, london, U.K., November 1996.

[45] F. Kelly. “Charging and Rate Control for Elastic Traffic”. *European Transactions on Telecommunications*, 8, 33–37, 1997.

[46] A. Kolarov and G. Ramamurthy. “A Control Theoretic Approach to the Design of Closed Loop Rate Based Flow Control for High Speed ATM Networks”. In *Proc. of the IEEE INFOCOM’97*, 3a.3, pages 293–301, 1997.

[47] A. Kolarov, G. Ramamurthy, T. Takamichi, and T. Murase. “Impact of Misbehaving Users and The Role of Policers in ABR Service”. In *Proc. of the GLOBECOM’98*, Sydney, Australia, November 1998.

[48] D. Kouvatsos, editor. *“ATM Networks Performance Modelling and Analysis”*, volume 2. Chaman & Hall, London, 1996.

[49] H.T. Kung. “The FCVC (Flow-Controlled Virtual Channels) Proposal for ATM Networks”. In *Proc. of the International Conf. on Network Protocols*, pages 116–127, San Franicso, California, October 1993.

[50] H.T. Kung and A. Chapman. “Credit-based flow control for ATM networks: Credit update protocol, adaptive credit allocation, and statistical multiplexing”. In *Proc. of the SIGCOMM’94*, volume 24, pages 101–114, October 1994.

[51] G. Latouche and V. Ramaswami. “A logarithmic reduction algorithm for Quasi-Birth-Death processes”. *Journal of Applied Prob.*, 30, 650–674, 1993.

[52] S. Low and P. Varaiya. “A New Approach to Service Provisioning in ATM Networks”. *IEEE/ACM Transactions on Networking*, 1(5), October 1993.

[53] J.k. MacKie-Mason and H.R. Varian. *“Some Economics of the Internet”*. University of Michigan, February 1994.
`http://www.sims.berkeley.edu/ hal/people/papers.html`

[54] J.k. MacKie-Mason and H.R. Varian. *“Some FAQs about Usage-Based Pricing”*. University of Michigan, August 1994.
`http://www.sims.berkeley.edu/ hal/people/papers.html`

[55] S. Manthorpe and J. Le Boudec. “A comparison of ABR and UBR to support TCP traffic”. Technical Report RT-97-224, EPFL, February 1997.

[56] K. Möttönen. “W-ERICA, Description of a simple and efficient ABR control algorithm and simulation results of TCP/IP over ABR and UBR in the presence of realistic VBR traffic”. Technical report, COST-257, May 1997.

[57] S. Muddu, F. Chiussi, C. Tryfonas, and V. Kumar. “Max-Min Rate Control Algorithm for Available Bit Rate Service in ATM Networks”. In *Proc. of the IFIP WG–6.2 1st Workshop on ATM Traffic Management WATM’95*, Paris, France, December 1995.

[58] J. Murphy, L. Murphy, and E.C. Posner. “Distributed Pricing for Embedded ATM Networks”. In *Proc. of the ITC 14*, J. Labetoulle and J.W. Roberts, editors, pages 1053–1063. Elsevier, 1994.
`http://www.eeng.dcu.ie/ murphyj/publ/publ.html`

[59] M.F. Neuts. “Markov Chains with Applications in Queueing Theory, which have a Matrix-Geometric Invariant Probability Vector”. *Journal of Applied Prob.*, 10, 185–212, 1978.

[60] M.F. Neuts. *"Matrix-Geometric Solutions in Stochastic Models"*. The John Hopkins University Press, Baltimore, 1981.

[61] M.F. Neuts. *"Structured Stochastic Matrices of M/G/1 Type and Their Applications"*. Marcel Dekker, New York, 1989.

[62] K.K. Ramakrishnan and P. Newman. "Integration of Rate and Credit Schemes for ATM Flow Control". *IEEE Network Magazine*, pages 49–56, 1995.

[63] M. Ritter. "Congestion Detection Methods and their Impact on the Performance of the ABR Flow Control". In *Proc. of the ITC 15*, V. Ramaswami and P.E. Wirth, editors, pages 1107–1117. Elsevier, June 1997.

[64] M. Ritter. "The Effect of Bottleneck Service Rate Variations on the Performance of the ABR Flow Control". In *Proc. of the IEEE INFOCOM'97*, 7a.4, pages 815–822, 1997.

[65] M. Ritter and P. Tran-Gia. "Performance Analysis of Cell Rate Monitoring Mechanisms in ATM Systems". In *Proc. of the International Conference on Local and Metropolitan Communication Systems*, pages 134–139, Kyoto, 1994.

[66] J. Roberts, U. Mocci, and J. Virtamo, editors. *"Broadband Network Teletraffic - Final Report of Action COST 242"*. Springer Verlag, 1996.

[67] J.W. Roberts. "What ATM Transfer Capabilites for the B-ISDN?". In *Proc. of the IFIP WG–6.2 1st Workshop on ATM Traffic Management WATM'95*, Paris, France, December 1995.

[68] L. Roberts. "Enhanced Proportional Rate Control Algorithm (EPRCA)". ATM Forum contribution number 94-0735R1, August 1994.

[69] A. Romanov and S. Floyd. "Dynamics of TCP Traffic over ATM Networks". *IEEE Journal on Selected Areas in Communications*, 13(4), 633–641, May 1995.

[70] H. Saito, M. Ishizuka, and K. Kawashima. "Available Bit Rate service: open issues". In *Proc. of the ITC Specialists Seminar on Control in Communications*, pages 165–176, Lund, Sweden, September 1996.

[71] R. Satyavolu, K. Duvedi, and S. Kalyanaraman. "Explicit rate control of TCP applications". ATM Forum contribution number 98-0152, February 1998.

[72] S. Shenker. "Fundamental Design Issues for the Future Internet". *IEEE Journal on Selected Areas in Communications*, 13(7), 1176–1188, September 1995.

[73] D. Songhurst and F. kelly. "Charging Schemes For Multiservice Networks". In *Proc. of the ITC 15*, V. Ramaswami and P.E. Wirth, editors. Elsevier, June 1997.

[74] W.R. Stevens. *"TCP/IP Illustrated, Volume 1: The Protocols"*. Addison–Wesley, 1994.

[75] W.R. Stevens. "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms". RFC 2001, January 1997.

[76] H. Suzuki and F.A. Tobagi. "Fast Bandwidth Reservation Scheme with multi-Link *&* Multi-Path Routing in ATM Networks". In *Proc. of the IEEE INFOCOM'92*, 10a.2, pages 2233–2240, 1992.

[77] D.H.K. Tsang and W.K.F. Wong. “A New Rate-Based Switch Algorithm for ABR Traffic to Achieve Max-Min Fairness with Analytical Approximation and Delay Adjustment”. In *Proc. of the IEEE INFOCOM’96*, 9d.1, pages 1174–1181, 1996.

[78] J. Witters, G.H. Petit, E. Metz, and E. Desmet. “Throughput Analysis of a DGCRA-based UPC function monitoring misbehaving ABR end-systems”. In *Proc. of the IEEE ATM’97 Workshop*, pages 204–213, Lisboa, May 1997.

[79] N. Yin. “Fairness Definition in the ABR Service Model”. ATM Forum contribution number 94-0928R2, November 1994.

[80] N. Yin and M. Hluchyj. “On Closed-Loop Rate Control for ATM Cell Relay Networks”. In *Proc. of the IEEE INFOCOM’94*, 1c.4, pages 99–108, 1994.