A new charging scheme for ATM based on QoS

Babul Miah
Supervisor: Professor Laurie Cuthbert

Submitted for the Degree of Doctor of Philosophy

Department of Electronic Engineering
Queen Mary and Westfield College
University of London
United Kingdom

June 1999

In memory of Poppy Miah

ABSTRACT

New services are emerging rapidly within the world of telecommunications. Charging strategies that were appropriate for individual transfer capabilities are no longer appropriate for an integrated broadband communications network. There is currently a range of technologies (such as cable television, telephony and narrow band ISDN) for the different services in use and a limited number of charging schemes are applicable for each of the underlying technologies irrespective of the services used over it. Difficulties arise when a wide range of services has to be supported on the same integrated technology such as asynchronous transfer mode (ATM); in such cases the type of service in use and the impact it has on the network becomes much more important. The subject of this thesis, therefore, is the charging strategies for integrated broadband communications networks. That is, *the identification of the requirements associated with ATM charging schemes and the proposal of a new approach to charging for ATM called the "quality of service based charging scheme"*.

Charging for ATM is influenced by three important components: the type and content of a service being offered; the type of customer using the services; and the traffic characteristics belonging to the application supporting the services. The first two issues will largely be dependent on the business and regulatory requirements of the operators. The last item, and an essential one for ATM, is the bridge between technology and business; how are the resources used by a service quantified? Charging that is based on resource usage at the network level was the prime focus of the research reported here.

With the proposed charging scheme, a distinction is first made between the four different ATM transfer capabilities that will support various services and the different quality of service requirements that may be applicable to each of them. Then, resources are distributed among buffers set-up to support the combination of these transfer capabilities and quality of services. The buffers are dimensioned according to the M/D/1/K and the ND/D/1 queuing analysis to determine the buffer efficiency and quality of service requirements. This dimensioning provides the basis for fixing the price per unit of resource and time. The actual resource used by a connection is based on the volume of cells transmitted or peak cell rate allocation in combination with traffic shapers if appropriate. Shapers are also dimensioned using the quality of service parameters. Since the buffer

efficiency is dependent on the quality of service requirements, users (customers) of ATM networks buy quality of service. The actual price of a connection is further subjected to a number of transformations based on the size of the resource purchased, the time of the day at which a connection is made, and the geographical locality of the destination switch.

It is demonstrated that the proposed charging scheme meets all the requirements of customers and of network operators. In addition the result of the comparison of the new scheme with a number of existing, prominent, ATM charging schemes is presented, showing that the performance of the proposed scheme is better in terms of meeting the expectations of both the customers and the network operators.

ACKNOWLEDGEMENT

Although the content of this thesis represents the sole contribution of the author, the research in to the much wider issues surrounding charging and billing involved number of people from whom came the guidance and encouragement. In recognition of this the author would like to thank his supervisor Professor Laurie Cuthbert and his mentor Professor John Griffiths; their support has been the real motivation for this work. The other members of the advanced telecommunications laboratory in QMW are Dr Jonathan Pitts, Dr John Schormans and Dr Eric Scharf, all whom have provided a lot of feedback and comments on all aspects of broadband networking including modelling, simulations and performance evaluation. The author offers his greatest appreciation to them all.

Finally, sincere thanks to all the member of the QMW Advanced Telecommunications Laboratory, Distributed Artificial Intelligence Laboratory, Antenna Laboratory, the departmental technicians and the secretaries for their source of information, helpful suggestions and endless entertainment during my term in QMW.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF MATHEMATICAL SYMBOLS

α	Total cost/tariff of overall switch capacity in unit bandwidth
β	Resource used by a connection
FC	Fixed cost for connection set-up or re-negotiation
SC	Subscription cost
VC	Variable cost
T	Duration of a connection
x_i	Load produced (bandwidth) at a single instant
P_{xi}	Probability of producing load x_i
E[X]	Mean load produced by a source
X	Representation of load produced ($X=x_i,\ldots,x_c$)
C	Switch capacity (link rate)
N	Maximum number of connections
n	Number of connections
Pr[cell_lost]	Probability of cells lost
z	Resource usage information in Kelly's charging scheme
a	Constant in Kelly's charging scheme to represent tariff parameter
	Co-efficient in Lindberger's charging scheme
b	Constant in Kelly's charging scheme to represent tariff parameter
	Co-efficient in Lindberger's charging scheme
B(z)	Effective bandwidth in Kelly's charging scheme
Charge	Net charge of connection (VC+FC+SC)
Price	Charge for resource usage (variable cost)
q	Cell loss probability parameter as in 10^{-q}
d	Effective bandwidth in Lindberger's charging scheme
k	Constant in Botvich's charging scheme
m	Mean bandwidth
eb	Equivalent bandwidth in Griffiths' charging scheme
dr	Design Rate in Griffiths' charging scheme
s	Constant (Kelly's effective bandwidth) to represent space parameter
e	Number 2.7183
R	Source rate distribution function

z_d	Declared resource usage information
z_m	Measured resource usage information
z_{mo}	Measured resource information which is above z_d
z_{mu}	Measured resource information which is below z_d
CLP	Cell loss probability in 10^{-q} (e.g., 10^{-8})
M(s)	Effective bandwidth formula
egress_C	Egress buffer service rate
ingress_C$_j$	Ingress buffer service rate for stream j
ingress_bd$_j$	Ingress bandwidth distribution factor for stream j
*egress_*ρ	Egress buffer efficiency factor
*ingress_*ρ_j	Ingress buffer efficiency factor for stream j
net_ingress_C$_j$	Net bandwidth available for stream j
ζ	QoS requirement (specified in CLP and CTD pair)
τ	Time of the day a connection is made
ω	Geographical location of the destination node
$\alpha(\zeta,\beta,\tau,\omega)$	Cost/Tariff as a function of ζ, β, τ, ω in QoS based charging scheme
Q(x)	Probability of number of cells in buffer exceeding x
egress_ctd	Maximum delay in egress buffer
ingress_ctd$_j$	Maximum delay in ingress buffer for stream j
lt	Lost traffic
rx	Offered traffic
tx	Carried traffic
E[a]	Average arrival
s(0)	Probability of zero cell in the buffer
a(k)	Probability of k arrival in the buffer
λ	Mean arrival rate
egress_b	Egress buffer size
ingress_b$_j$	Ingress buffer size for stream j
B	Probability of call blocking
shaper_b$_j$	Shaper buffer size for stream j
slr$_j$	Shaper leak rate for stream j
A	Offered traffic used in Erlang's equation

y	Constant (real number) for shaper dimension
$egress\_\alpha$	Total cost/tariff of egress service rate in unit bandwidth
$ingress\_\alpha_j$"	Percentage of $egress\_\alpha$ distributed to stream j
$ingress\_\alpha_j$'	Total cost/tariff of ingress buffer rate in unit bandwidth for stream j
$ingress\_\alpha_j$	Total cost/tariff of ingress buffer rate per unit bandwidth per unit time for stream j
ctd_sw_j"	Total delay in the switch due to ingress and egress buffers
ctd_sw_j'	Delay allowed by the switch given the CTD
ctd_sw_j	Delay allowed by the switch given the CTD and the transmission delay

GLOSSARY

Service	End products used by the customers.
Applications	Software and hardware tools required supporting a service.
Source	Combination of a particular service and application represented by a source generating traffic as a stream of ATM cells.
Connection	A single connection reaching from end to end.
Stream	A stream of ATM cells on a connection or link.
Provider	Organisation responsible for provisions of a service which end user may purchase.
End-User	Individuals or organisations who use the services.
Customer	Individuals or organisations that pays for the uses the ATM services. This could be either an end user or a service provider.
Operator	Organisation responsible for networks provisioning to carry services from service providers to end-users. Both the end users and the service providers are customers of network operators.
AAL	ATM Adaptation Layer
ABR	Available Bit Rate
ABT	ATM Block Transfer
ACTS	Advanced Communications Technologies and Services
ATC	ATM Transfer Capability
ATM	Asynchronous Transfer Mode
BISDN	Broadband Integrated Services Digital Network
CAC	Connection Admission Control
CATV	Cable Television
CBR	Constant Bit Rate or Continuous Bit Rate
CER	Cell Error Ratio
CC	Cross-Connection
CDV	Cell Delay Variation
CLP	Cell Loss Priority
CTD	Cell Transfer Delay
CSEC	Charging Scheme Evaluation Criteria
DBR	Deterministic Bit Rate

DT	Delayed Transmission
EBW	Effective Bandwidth
FTP	File Transfer Protocol
GCRA	Generic Cell Rate Algorithm
GoS	Grade of Service
HEC	Header Error Check
IP	Internet Protocol
ISDN	Integrated Services Digital Network
IT	Immediate Transmission
ITU-T	International Telecommunications Union Telecommunications
MCR	Minimum Cell Rate
MIB	Management Information Base
NASP	Network Service Access Point
NE	Network Element
NEC	Network Element Component
NEF	Network Element Function
NMS	Network Management System
NO	Network Operator
NPD	Network Provider
NPR	Network Parameter Control
nrt	Non-Real Time
OAM	Operations and Maintenance
PCR	Peak Cell Rate
POTS	Plain Old Telephone System
PVC	Permanent Virtual Connection
QoS	Quality of Service
QBCS	Quality of Service Based Charging Scheme
rt	Real Time
SBR	Statistical Bit Rate
SCR	Sustainable Cell Rate
SVC	Switched Virtual Circuit
TMN	Telecommunications Management Network
UBR	Unspecified Bit Rate

UNI	User Network Interface
UPC	Usage Parameter Control
VBR	Variable Bit Rate
VC	Virtual Channel
VCC	Virtual Channel Connection
VP	Virtual Path
VPC	Virtual Path Connection
VPL	Virtual Path Link
WFQ	Weighted Fair Queuing

1. INTRODUCTION

Currently a range of technologies exists for varying services: technologies such as POTS for telephony, CATV for television or video, ISDN for data transfer and IP for information sharing. ATM can enable the convergence of all of these technologies and increase the capability of the network to support different services. For example, a user who requires a video on a Saturday night to view at home currently visits their local video shop and physically carries it home for viewing. However, ATM has the potential to bring users (the viewer), the providers (the video shop) and the operators (the carriers) together at a "press of a button". In the long term, ATM will also reduce the cost of service provisioning through concentration of efforts in one area of development; data, telephony, television and alike, will all be served by the single technology.

Many advances have already been made in technology. The difficulty for the operators is to convince the users and the providers of the services that ATM is what they need. In the real world, this essentially means convincing them that it makes good economic sense; this raises the question, what will it cost [IRV01]? In order to answer this question, the operators must deduce the cost and produce a charging strategy that will suit the requirements of all the parties involved. Finding this strategy is, however, a challenge. For example, if POTS, ISDN and IP are provided over a single link, should the charge be based on the rates for POTS, ISDN and IP [GAB01][SRI01]? If so, there is little incentive for replacing current infrastructure and converting to ATM. On the other hand, charging on a single basis (such as counting the number of cells transmitted) regardless of the service in use will not be adequate as this will make some services more expensive while making others far too cheap compared to their impact on the network as a whole. For example, two identical services, using different ATM transfer capabilities (such as CBR and VBR) may transmit an equal number of ATM cells but the impact on the network (in terms of resource allocation) by the two connections supporting the services may differ hugely. Clearly, if these difficulties are not overcome the demand for ATM will not occur, resulting in still higher costs for the technology, pushing users and providers even further from accepting this revolution in communications.

It is crucial, if ATM is to survive in the market place, to identify the charging requirements of all the major players in the telecommunications sector and develop a suitable charging scheme.

1.1. CHARGING SCHEMES

Charging for ATM will be influenced by the following key factors:

- Calculation of the network resource usage information (not just simply time and volume).
- Identification of the service type (e.g., telephony or video on demand) and value of the service (e.g., *latest release* of a video).
- Identification of the customer group (e.g., residential or business users requiring network resources from network operators to deliver a service).

The charging scheme itself is a mathematical expression and is a function of the charging parameters (dependent on the resource usage) and tariff parameters (dependent on service type and customer group).

The argument for linking the charging scheme to the measure of resource usage is one of fairness: customers should pay for what they have used. In the world of perfect competition the market will drive charges ever closer to just the recovery of the cost of the network. Resource usage based charging schemes give flexibility to both the customers and the network operators.

1.1.1. Network Resource Usage

A particular application can support a range of service types, while a particular service may use a range of applications. Each case can be represented by a traffic source emitting a stream of traffic (ATM cells) that will have a varying degree of impact on the network. Identification of the actual resource used by a particular connection is one way to ensure that a charging scheme is fair and flexible.

1.1.2. Services and Customers

As in all businesses, survival in a competitive market will require more than just fairness. For example, consider the cost of transporting a business class customer and an economy

class customer by an airline (assuming the seating density is similar, as on short-haul flights). The difference in cost is small compared to the difference in the actual charge the two customers will pay; one is effectively subsidising the other. Hence, it is important to realise how realistic a charging scheme is for any particular customer and service pair, and what benefit that service brings to the customer.

1.2. OBJECTIVES OF THE THESIS

The above description served to highlight the importance of developing a suitable charging scheme for ATM. The aim of the research reported in this thesis has been to:

- *determine the requirements for ATM charging schemes;*
- *develop a charging scheme for ATM networks that meets these requirements.*

The focus is the resource usage information discussed in section 1.1.1. and development of a charging scheme based on the resources used by a service.

In order to meet the above objectives, the initial phase of the research concentrated on the impact that ATM will have on charging. A description of charging in general and the related issues is presented in the thesis. In particular, the author describes how customers view services that are run over ATM networks and the need for this to be mapped onto network related parameters (used by the charging schemes). One aspect of customer perception of services that has been identified as significant, and which led to the development of the proposed charging scheme, is the issue of quality of service (QoS). An approach to QoS mapping (between customer's perception and network parameters) is therefore also presented in the thesis.

For the presentation of the main thrust of the research, the author first *proposes a set of rules for evaluating a charging scheme. This is a list of charging scheme evaluation criteria*, based on which a qualitative evaluation of a number of current resource usage based charging schemes, widely published in the literature, is carried out. Preliminary evaluation of the existing charging schemes led to the design of a new charging scheme that is based on the quality of service available with the ATM technology. Description of this new scheme is then presented in the thesis. The proposed and the existing schemes were then put to the test using simulations, which provided the necessary information for

experimental evaluation of the charging schemes. The results of this evaluation are also described here.

With the *new approach, called the QoS based charging scheme*, a set of QoS streams is made available to the customers. Each stream is dimensioned to support a particular QoS. The cost (to the customer) of each of the QoS streams is different, is dependent on the QoS, and is announced as price per unit of resource usage per unit time.

Customers, knowing the price in advance, choose a particular stream according to their QoS requirements and declare the resource needed. The resource requirement, β, is specified in terms of $y*m$ (or minimum cell rate for ABR and peak cell rate for CBR based services). The parameter m is the mean rate of the traffic source representing the desired service and y is a constant (varying from 1 to 5). The parameter y allows adjustment of the bandwidth requirement to further improve the quality or reduce price. For the UBR based services (which require no QoS guarantee), the bandwidth declaration is not necessary; charges are made based on the volume of cells transmitted.

The price of usage of a QoS stream j depends on the cost function $\alpha_j(\zeta,\beta,\tau,\omega)$, which is a function of:

1. quality of service parameters, ζ;
2. size of the resource purchased, β;
3. time of the day the connection is made, τ;
4. the geographical location of the destination switch, ω.

The QoS streams are dimensioned with a set of buffers, including a traffic shaper for VBR traffic. Dimensioning is performed using the M/D/1/K and ND/D/1 queueing analysis.

1.3. New Contributions from the Thesis

Those sections of the thesis where *the author has made new contributions are:*

1. *Proposal of the charging scheme evaluation criteria – section 3.4.*
2. *Qualitative evaluation of the existing charging schemes – section 3.6.*
3. *Proposal of the quality of service based charging scheme – chapter 4.*
4. *Simulation, simulation analysis and experimental evaluation of all the charging schemes – chapter 5.*

Item (1) was presented by the author as work input to the European ACTS project CANCAN (AC014), set-up to look at charging for ATM networks. The author was a participant of this project, which ended in September 1998. However, the work presented in this thesis is the sole contribution of the author except that specifically referenced to others.

The approach taken in design of the proposed QoS based charging scheme has to make charging an integral part of network configuration, dimensioning and management.

1.4. Structure of the Thesis

A list of acronyms and terms is provided in the glossary defining, for example, the difference between "user" and "customer". In addition, figures, tables, and mathematical symbols used throughout the thesis are also listed at the beginning. The main body of the report consists of seven chapters including the introduction, the discussion and the conclusion. Each chapter begins with a brief description of its scope/objectives and ends with a brief summary of the outcome. Details about each chapter are provided below.

Chapter two provides a background into ATM, focusing only on components significant in deployment of a charging scheme.

In chapter three, a background into charging for ATM is provided, including the description of a general charging scheme algorithm and the quality of service issues related to charging for ATM. In particular, a mechanism for QoS mapping between different layers of the OSI protocol stack is put forward. Also presented in chapter three is the proposed charging scheme evaluation criteria and the description of the most prominent charging schemes in the literature. The chapter concludes with a presentation of the qualitative evaluation and comparison of the existing charging schemes based on the proposed criteria.

In chapter four, the new quality of service based charging scheme is described in detail. Simulations performed to evaluate and compare the proposed and existing charging schemes are presented in chapter five.

Discussion on the topics covered in the thesis and the conclusions reached from the discussion are presented in chapter six and seven.

References and Appendixes are provided in the remaining chapters.

2. ATM TECHNOLOGY

2.1. OVERVIEW

ATM has been identified as the transfer mode for the broadband integrated service digital network (BISDN) by the ITU-T, and is capable of supporting all current and anticipated future services, while maintaining an efficient use of the available network resources. It is a merger between different transport capabilities over a unique interface (I/F in Figure 1) to a single integrated network to handle the entire customer's needs; see Figure 1. The technology is transparent to the applications being run by the customers, residing only in the lower layers of the OSI reference model.

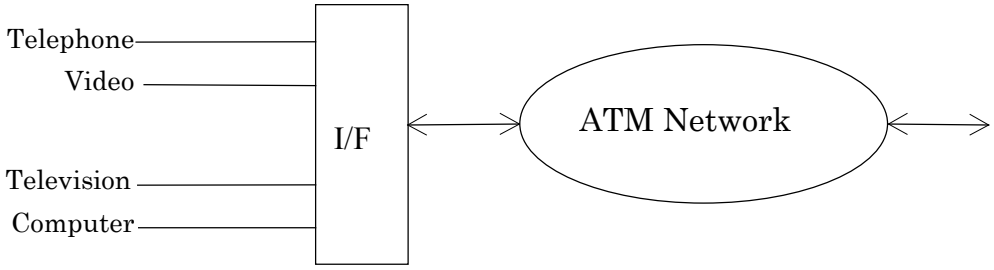

Figure 1: Simple ATM network architecture.

Figure 2 shows the ATM reference model [CUTH01]. The control plane, user plane, management plane, ATM adaptation layer and the ATM layer influence the charging mechanism. The management plane is responsible for collection of charging parameters from the lower layers as well as the service and user information from the user plane to process the bill. The management plane can also communicate with the control plane to control the network to optimise network usage in terms of utilisation and profitability.

Since ATM has already been described in detail in the literature, for example [CUTH01], only a brief outline of the principle of it is given here, only including those components that are significant in deployment of a charging scheme:

1. Key network functions and network elements
    - user parameter control
    - traffic shaper
    - connection admission control
    - switching functions

2. ATM layer consideration
    - traffic descriptor
    - ATM transfer capabilities
    - quality of service
    - traffic parameters
3. ATM adaptation layer consideration
4. Higher layer service and applications.

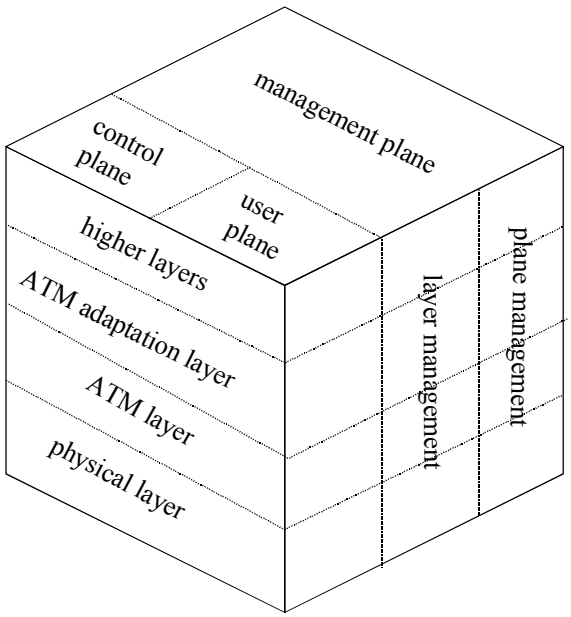

Figure 2: ATM reference model.

2.2. KEY NETWORK FUNCTIONS AND NETWORK ELEMENTS

To ensure the desired broadband network performance, an ATM based network will have to provide a set of traffic control capabilities. Some possible solutions are provided in [ITU01]. Below are the list of four that are important for the charging scheme under consideration:

- user parameter control (UPC);
- traffic shaper;
- connection admission control (CAC);
- switching and multiplexing.

2.2.1. User Parameter Control

UPC is a set of actions taken by the network to monitor and control the traffic flow in a network. ITU-T specifies [ITU01] that UPC is required at the public UNI in order to ensure that parameters agreed upon in the traffic contract are not violated by the customers. UPC may be applied to VCCs or VPCs. Mainly the following traffic parameters are negotiated during the connection set-up phase and can be enforced by an UPC function:

- the peak cell rate (PCR) and the corresponding cell delay variation tolerance;
- the sustainable cell rate (SCR), burst tolerance (BT), and cell delay variation of bursty traffic.

How these parameters are controlled by the UPC is network specific. However, in real implementations, most UPC functions use the conformance definitions for the corresponding parameters that are based on the generic cell rate algorithm (GCRA) specified in [ITU01].

According to the GCRA if a cell is detected to be non-conforming the UPC function can perform one of the following actions:

- cell passing;
- cell tagging;
- cell discarding.

Tagging means that the cell loss priority bit in the cell header is set to “1”. In real systems, cell discarding is the easiest to implement, particularly as it is not guaranteed that a subsequent node recognises CLP differences.

2.2.2. Traffic Shaper

While UPC is intended for monitoring and policing traffic, traffic shaping is defined to be a mechanism that alters the traffic characteristics of a stream of cells on a connection to achieve better network efficiency whilst meeting the QoS objectives, or to ensure conformance at a subsequent interface.

Traffic shaping can be applied to VCCs or VPCs. Different objectives of shaping are possible:

- limitation of cell rate to the peak cell rate PCR;

- reduction of burst size;
- limitation of average cell rate within an interval Δ to the sustainable cell rate (SCR);
- reduction of CDV.

The point where traffic shaping is carried out has not been fully defined. Within the network it may be used to get a smoother traffic flow that needs less resources inside the switching nodes. However, the main application of traffic shaping is cell spacing at the customer side. In this case the purpose of shaping is to ensure that the cells generated by the source are conforming (e.g., in relation to the single rate shaper) to the negotiated traffic contract before they are transmitted over the user network interface (UNI).

2.2.3. Connection Admission Control

One of the fundamental issues in ATM networking is connection admission control (CAC). When a new connection is requested by a customer, the network must decide whether or not to admit the connection; and if so, how to route it through the network and what resources (bandwidth) to reserve for its virtual channel. Packet-switched networks use higher-layer protocols to guarantee acceptable packet delivery but these are not expected to scale well to broadband speeds. In circuit-switched networks (such as most telephone networks) the CAC mechanism results in connection blocking when the bandwidth of a requested connection exceeds the available bandwidth. But in an ATM network the traffic source may be bursty, so the required bandwidth of its virtual channel varies with time during the connection.

The nature of this time-varying bandwidth differs widely among different sources and therefore it is difficult to characterise the bandwidth requirement of a connection. This difficulty has led to proposals to reserve the peak bandwidth of the connection (deterministic multiplexing), as required for constant bit rate (CBR) sources. However, the gain in efficiency possible by taking advantage of the statistical nature of variable bit rate (VBR) sources has led to many schemes for statistical multiplexing. Such schemes assign less than the peak bandwidth requirement, and therefore may introduce cell loss and/or delay. The extent to which the service degradations occur due to such action is measured by the quality of service (QoS) offered by the network.

One mechanism for the CAC that has been widely researched is the notion of effective bandwidth instead of peak bandwidth for resource allocation [MUR01] [COST01] [HUI01] [KELL01] [LIND01]. Effective bandwidth is a way of summarising the statistical information of a source in a single parameter. The complex problem of resource allocation of a multi-service network can be simplified by trying to get an equivalent circuit switched model. By using effective bandwidth it is possible to get a linear equation (similar to the circuit switched networks) and see if there is sufficient bandwidth left to admit another connection. Algorithms for calculating effective bandwidth calculation have been developed in numerous literatures, the most prominent has been the one developed by [HUI01].

2.2.4. Switching Function

Network elements have switching functions and are defined as components necessary to transport a cell from the source to its destination (sink). The following types of network elements can be distinguished (see Figure 3):

- link;
- switch;
- multiplexer/de-multiplexer.

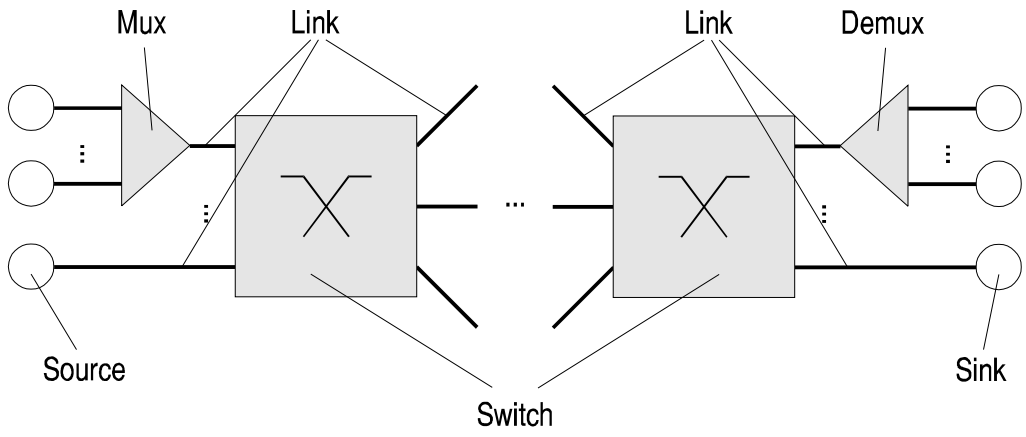

Figure 3: Typical network elements.

A source is connected to a multiplexer or a switch via a link. On the destination side of the network a sink is connected to the output of a de-multiplexer or a switch. A de-multiplexer could also be seen as a special kind of switching node providing merely the switching functionality without buffering. Therefore, it does not have to be modelled as a separate

network element but rather as a simplified derivation of a switch model. In a similar way a multiplexer is also a special kind of switch with buffering functionality to avoid collisions.

2.3. ATM Layer Consideration

ATM is a transfer mode that transmits customer information in fixed size packets called cells. Cells are 53 octets long, of which up to 48 octets represents the actual customer information (the payload) and 5 octets cell header which contains the network information such as routing and source destination addresses. The actual size of the payload depends on the type of service being used as number of additional octets may be set a side for application related control information: the AAL information.

Due to the high quality of modern transmission systems and the fast data rate, no provision for error recovery on the payload is performed; flow control and error control are carried out on an end-to-end basis at a higher layer, except the ABR transfer capability. ABR transfer capability makes provision for flow control within the networks. Only a limited number of header checks are performed on the node to node basis.

ATM is connection oriented, with cells associated with a number of connections being multiplexed onto a single link using an asynchronous time division multiplexing technique. The transmission rates of the applications are independent of the transmission rate of the network and the source can generate and transmit cells asynchronously. If a source does not generate cells, these empty slots are available for other use. Cells transmission between nodes is asynchronous; synchronisation is maintained by the transmission system. Line coding is used to maintain clock synchronisation and cell delineation is performed using a procedure based on the error control field (HEC) of the cells [CUTH01].

The connection-oriented nature of ATM is purely logical using virtual channel identifiers (VCI) and virtual path identifiers (VPI) to identify connections on a link. The connections preserve cell sequence integrity: cells on a particular connection will always arrive at their destination in the same order that they were sent.

2.3.1. Traffic Descriptor

A customer requests a connection for a specific service by specifying a traffic descriptor. The connection is then established by the connection admission control (CAC) function in the control plane based on the resources required for that service. The traffic descriptor is an identification of the transfer capability, quality of service (QoS) parameters and traffic parameters.

When a connection has been set-up, an UPC may also be set-up to ensure that the customer does not violate the negotiated traffic descriptor values. In addition, a traffic shaper may be put in place to minimise the negative impact of the service on the network, a fact that can have an implication on charging.

Transfer capability, QoS and traffic parameters together describes a traffic source emitting traffic (cells) on a connection such that characteristics of the traffic stream on the connection can be analysed to compute a charge.

2.3.2. ATM Transfer Capabilities

Transfer capability defines the exact way in which ATM cells are transmitted from the customer terminal. Currently several different capabilities are in existence. Those from the ITU-T specifications are shown below [ITU01]:

- deterministic bit rate (DBR);
- statistical bit rate (SBR);
- ATM block transfer with delayed transmission (ABT/DT);
- ATM block transfer with immediate transmission (ABT/IT);
- unspecified bit rate (UBR).

Likewise, those from the ATM Forum are [ATMF01]:

- constant bit rate (CBR);
- real time variable bit rate (rt-VBR);
- non real time variable bit rate (nrt-VBR);
- unspecified bit rate (UBR);
- available bit rate (ABR).

A full description of each of these can be found in [ITU01][ATMF01]. Within the context of this work, the author considers the ATM Forum specifications for the design of the proposed charging scheme; i.e., CBR, VBR (real time and non real time), UBR and ABR capabilities.

2.3.3. Quality of Service

QoS parameters define a set of constraints for the connection, with which the network must conform in order to maintain a desired level of service quality. There are three level of QoS: those at the service level, such as flickers on the video screen and clicks on the telephone; those at the application level, such as error control and flow control; those at the network level, such as cell loss probability and cell delay. QoS at the service and application level must be mapped onto the QoS at the network level so that appropriate resources may be allocated to meet these criteria. The following may be negotiated between the end systems and the network:

- peak to peak cell delay variation (CDV);
- maximum cell transfer delay (CTD);
- cell loss probability (CLP).

In addition, the following parameters are identified and may also be negotiated:

- cell error ratio (CER);
- severely error cell block ratio;
- cell miss-insertion ratio.

An initial set of classifications have been made, by the ITU-T [ITU01], of the QoS values and the possible services that may require this constraint, see Table 1. These relate to the delay and the cell loss probability of traffic streams belonging to a general set of service type (e.g., multimedia services). The parameters in this classification are derived from the technical capability of the network versus the connection.

Class	Rate	Cell Loss Ratio	Maximum Delay	Description	Example Traffic
0	unspecified	unspecified	unspecified	best effort, UBR or ABR	file transfer, email
1	Specified	1E-8	1 ms	circuit emulation, CBR	digital private line
2	Specified	1E-7	10 ms	VBR, large burst size	multimedia, video
3	Specified	1E-4	1000 ms	connection oriented data	frame relay
4	Specified	1E-4	100 ms	connection less data	IP and SMDS

Table 1: ITU-T QoS classifications.

2.3.4. The Traffic Parameters

The traffic parameters are:

- peak cell rate (PCR);
- mean cell rate (MCR);
- sustainable cell rate (SCR);
- maximum burst size (MBS);

Table 2 shows a typical relationship between the traffic parameters, QoS parameters and transfer capabilities according to the ATM-F specifications [ATMF01].

	CBR	rt-VBR	nrt-VBR	UBR	ABR
PCR	Specified	Specified	Specified	Optional	Specified
SCR/MBS	N/A	Specified	Specified	N/A	N/A
MCR	N/A	N/A	N/A	N/A	Specified
CDV	Specified	Specified	Unspecified	Unspecified	Unspecified
CTD	Specified	Specified	Unspecified	Unspecified	Unspecified
CLP	Specified	Specified	Specified	Unspecified	Varies

Table 2: Mapping transfer capabilities and QoS parameters.

2.4. ATM ADAPTATION LAYER CONSIDERATION

ATM network, the part of the network which processes the functions of the ATM layer, is independent of the telecommunications services it carries. This means that the customer payload is carried transparently by the ATM network and the ATM network does not process the customer payload and does not know the structure of the data units. This is known as semantic independence. There is also time independence as there is no timing

relationship between the clock of the application and the clock of the network, the network having to cope with any application bit rate.

The ATM adaptation layer (AAL) performs the necessary mapping between the ATM layer and the higher layers. This is done in the terminal equipment or terminal adapter, i.e., at the edge of the ATM network.

2.5. Higher Layer Services and Applications

Service at the highest layer is the end product used by the customers. For example, watching a video or participating at a videoconference. The applications that support these services are the software and hardware components that drive the physical equipment such as the monitor, the camera, the telephone and the editing tools.

It is important to identify these two higher layer components (services and applications) since the traffic descriptor required will be greatly affected by them.

2.6. Summary

In this chapter, a brief look at the ATM networking techniques and their functionality was provided. These are the key components that will enable different services to run over an ATM network and thereby effect the way a network operates and performs. Not surprisingly, what has been discussed in this chapter forms the basis for the charging scheme that is developed in the later chapters. Throughout the remainder of the thesis reference will be made to items described in this chapter; for example, traffic shaper, traffic descriptor, traffic parameter and quality of service.

The next chapter provides:

1. a background into the charging scheme in general and the associated issues;
2. the importance of QoS and a mapping strategy for QoS existing at different layers of the protocol stack, starting from customer perception of services to ATM layer QoS parameters such as CLP and CDV;
3. the list of charging scheme evaluation criteria;
4. the description of the key existing charging schemes;

Chapter three also provides a qualitative evaluation of the described existing charging schemes based on the criteria identified.

3. ATM CHARGING SCHEMES

3.1. OVERVIEW

Charging schemes are mathematical expressions that manipulate charging parameters for the purpose of producing bills. The most important part of a charging scheme from the technical point of view is the estimation of the resource used by a connection. For example, the expression

$$Charge = \alpha * \beta * T + FC + SC_i \qquad \text{(Equ 1)}$$

represents the total charge for a connection using resource, β, lasting a duration of T seconds. Here β and T are the charging parameters representing the total resource used by a connection. The constant α is the tariff parameter, which represents the cost (to the customer) of unit of resource used per unit time. The parameter FC is the fixed cost and is normally attributed to connection set-up or cost of negotiation, e.g., the switch resource used during connection set-up. In addition, the total charge may include a subscription cost (SC_i). This is the additional charge for facilities (such as connection holding and automatic answering) that a typical network operator might offer to its customers. The suffix *"i"* represents the service identity since there may be number of subscription costs included in a bill.

The previous chapter provided the background information relating to ATM technology. This chapter introduces further background information relating to charging for ATM network usage. The chapter then goes on to identify the importance of QoS and a strategy for mapping the QoS components that exists at different layers of the OSI protocol stack onto the ATM layer QoS parameters (used by the networks and the charging schemes).

Finally, a set of charging scheme evaluation criteria is proposed and four of the most prominent charging schemes in the literature are described. A qualitative evaluation of these schemes based on the proposed criteria is also presented.

3.2. GENERAL CHARGING SCHEME

Charging for ATM will in general consist of two main parts:

- Service related cost,

- type of service
- content of the service
- type of customer

- Subscription cost, SC_i,
  - additional network services
  - type of physical access
  - maximum bandwidth allowed
  - maximum number of simultaneous connections allowed
  - type of network elements (switches and gateways) involved
- Network usage cost, $\alpha*\beta*T+FC$,
  - Fixed part *(FC)*
  - Variable part ($\alpha*\beta*T$)

The variable part of the network usage cost is based on the resource used by a service associated with a connection and uses the charging parameters, β and T. β is dependent on the network elements, network functionality and traffic characteristic of the service in use (see section 1.1.1.). α is dependent on the service type, customer type and value of the service to the customer (see section 1.1.2.).

Let the variable part of the cost be expressed as the price of resource usage, in terms of:

$$Price = \alpha * \beta * T \qquad \text{(Equ 2)}$$

From a technical point of view the variable charge represents the greatest challenge and is the basis for development of the existing and the proposed resource usage based charging schemes discussed in this thesis.

3.3. QUALITY OF SERVICE

[GAR01] shows how ATM architecture can facilitate choices in service provisioning and QoSs. The new charging scheme presented later in the thesis is called the QoS based charging scheme. As the title suggests the scheme is based on customers paying for the quality of a service run over ATM networks. Research carried out by the European ACTS project CANCAN (AC014) concluded that most customers consider overall quality of service as the most important consideration when choosing a telecommunications supplier.

3.3.1. QoS at Different Layers

QoS as perceived by the customer is not directly identical to the QoS perceived by the network. In order to obtain a mapping of QoS at different levels, one needs to examine QoS in all other layers in the protocol stack. Table 3 gives details of the QoS that may be observed at the different layers.

Protocol Layer	QoS Requirements
Service	Perceived audio delay Perceived quality of visual pictures Files transfer delay Files lost Percentage of commitments met Customer reported faults per 100 lines Percentage of faults cleared to individual target lines Percentage of complaints resolved in 15 days Number of complaints about billing accuracy
Application	Fraction of cells suffering cell loss Fraction of cells suffering large delay Files corrupted
ATM	CDV CTD CLP
Physical	Peak signal to noise ratio System failure System reliability

Table 3: QoS at different level of protocol stack.

3.3.2. Relating QoS to Charging

QoS will affect charging in two ways. Firstly, the QoS at the ATM layer will directly influence the route a connection traverses, and thereby influence the charging parameters. However, QoS at the service and other layers will also affect charging through alteration of the tariff parameters or the charging parameters. For example, customers requiring equipment failure to be repaired within 24 hours may pay more than those who require it to be corrected within a week. Table 4 gives an indication of how QoS at different layers affects charging. The QoS requirements in any of the layers indicated in the second column either maps onto requirements in another layer or maps directly onto a charging parameter

(β) or a tariff parameter (α) as indicated in column 3. All requirements, however, will eventually map onto a charging or tariff parameter.

Protocol Layer	QoS Requirements	Mapped On To
Service	Audio quality	Map to all application layer QoS
	Pictures quality	Map to all application layer QoS
	File transfer delay	Map to all application layer QoS
	Files lost	Map to all application layer QoS
	Commitments met	*Map to Tariff Parameter*
	Faults per 100 lines	*Map to Tariff Parameter*
	Faults cleared	*Map to Tariff Parameter*
	Complaints resolved	*Map to Tariff Parameter*
	Billing accuracy	*Map to Tariff Parameter*
Application	Cell lost	Map to ATM layer CLP
	Cell delayed	Map to ATM layer CTD
	File corrupted	Map to ATM layer CLP
Physical	Signal noise	*Map to Tariff Parameter*
	System failure	*Map to Tariff Parameter*
	System reliability	*Map to Tariff Parameter*
ATM	CTD	*Map to Charging Parameter*
	CDV	*Map to Charging Parameter*
	CLP	*Map to Charging Parameter*

Table 4: Relationship between QoS and charging.

3.3.3. QoS Mapping Process

QoS perceived by a customer is a subjective measure of the quality and thus it is difficult to specify strictly how the service performance maps to QoS parameters. The following is one approach to QoS mapping proposed in [CC02].

Let *E[X]* be the mean load (average cell rate) produced by a connection, where

$$\mathrm{E}[X] = \sum_{i=1}^{C} x_i * P_{xi} \; ; \mathrm{X}=\mathrm{x}_1,\ldots,\mathrm{x}_\mathrm{C} \qquad \text{(Equ 3)}$$

C is the link capacity and x_i is the load (cell rate) and P_{xi} is the probability of the source producing load x_i. Then, over the time period T (considered to be the connection holding time), the total number of cells produced is N where

$$N = E[X] * T \qquad \text{(Equ 4)}$$

Let also q be the QoS parameter such that the cell loss probability, CLP, is

$$CLP = 10^{-q} \quad \text{(Equ 5)}$$

Then,

$$\Pr\{\text{sin}\, gle\_cell\_not\_lost\} = (1 - CLP) \quad \text{(Equ 6)}$$

$$\Pr\{cell\_not\_lost\_in\_connection\} = (1 - CLP)^N \quad \text{(Equ 7)}$$

$$\Pr\{connection\_affected\_by\_cell\_loss\} = 1 - (1 - CLP)^N \quad \text{(Equ 8)}$$

Applying Equ 8, three examples of traffic types for voice, video and data are provided in order to establish how ATM layer network performance affects user-perceived QoS.

Voice connections

Let the cell loss probability (CLP) be approximately 10^{-9} ($q=9$). This includes all cells whose delay is above an acceptable threshold. The CLP gives network performance, which may further translate into QoS, using the assumptions that a typical connection has the mean holding time of 3 minutes and a peak cell rate of 64kbit/s.

Using Equ 8 it can be deduced that a domestic customer, making and receiving one connection a day, is affected by cell loss once every 45 years; and a corporate customer making and receiving 500 connections a day being affected once every 2 months.

Video connections

The above procedure can be repeated for a real-time video link, using the assumptions that a typical connection has the mean holding time of 1 hour and a mean cell rate of 1Mbit/s.

Similar calculations give: a customer downloading one video a week will be affected by cell loss once every 99 weeks.

Data connections

It is now assumed that:

- all loss produces delay;
- packet size is 64 ATM cells;

- in the event of an error packet, re-transmission is required;
- total of 100Mbytes of data is to be transmitted.

With a CLP of 10^{-9}, a customer sending 100Mbytes of data will have a probability that no packets are retransmitted of 99.97%. With a CLP of 10^{-3}, a customer sending 100Mbytes of data will have to make 3 re-transmissions, comprising 2150 cells on average.

3.4. PROPOSAL: CHARGING SCHEME EVALUATION CRITERIA

In order to develop and evaluate an effective charging scheme, one needs to identify a set of criteria defining requirements to which it must conform in order to ensure that it fulfils the needs of both the customers and the operators. The author assumes that the customers require: a charging scheme that is simple to understand; be accountable; predictable; and enable choices in their use of services. Likewise, the operator's expectations are for it to be: simple to implement; enable control of the network operations; produce a steady flow of revenue; allow good selling points; and be secure. Taking these as the primary requirements a set of criteria can be deduced, which may be used to design and evaluate charging schemes. The criteria identified are listed below.

1. Clarity: How much does the customer need to know about the charging scheme and how understandable is the contract negotiation process? For example, must the customer declare complex traffic parameters or detailed traffic characteristics, the complexity of which may not commensurate with their mathematical knowledge?
2. Accountability: Can the customer's actual usage of the ATM service be traced in response to an audit query? The ease with which this can be done will depend on the extent and accuracy to which a connection is monitored. It will also depend on how well the procedure for calculating the charge is known and understood by the customers.
3. Predictability: Is the tariff known before a connection is made? It may not be if it depends on statistically varying properties of the applications supporting the services. How easy is it to predict the long-term charge? This could change if the charging scheme involves a periodic re-negotiation process with the network operator. Most customers appear to want a charging scheme that offers "no surprises".

4. Flexibility: Can the charging scheme be extended or modified to accommodate the charging aspiration and requirements of the operators? Can the charges be carried out easily and in a way that is understandable, fair and equitable to all interested parties?
5. Practicality: How easy and obvious is the implementation of the charging scheme. For example, does it depend on the CAC, and UPC and traffic shaper functions? Does it require online or offline measurements and does the accuracy of the measurement and calculation have any impact on the charge?
6. Control: In a competitive environment the charging scheme is an important factor in the operability of a network. It alone will determine how the network is utilised and the ability of the operator to recover costs and stay in business. Therefore charging schemes must help controllability of the network.
7. Choice: Does the charging scheme offer real choices to the customers in terms of different modes of operation according to the customers willingness to pay more or less? For example, can the customer select a low quality of service and get a cheaper deal for the same service?

Set against these criteria four existing charging schemes already proposed in the literature were investigated, followed by the proposal of an approach that capitalises the fundamentals of all of these schemes to derive a new mechanism for charging.

3.5. Existing Charging Schemes

Four of the most widely published charging schemes are:

1. the "effective bandwidth" scheme proposed by Frank Kelly [KELL01];
2. the "effective bandwidth" scheme proposed by Karl Lindberger [LIND01];
3. the "mean bandwidth" scheme proposed by Dmitri Botvich [CC01];
4. the "ATM filter" scheme proposed by John Griffiths [GRIF01][GRIF02].

3.5.1. Kelly's Charging Scheme

Kelly states that the price of a connection is [KELL01]:

$$Price = (a + bz) * T \qquad \text{(Equ 9)}$$

where a and b are constants set by the network operator and specifies the penalty the customer will incur if it does not comply with the bandwidth requirements specified during the connection set-up. T is the duration of the connection and z is the resource usage information defined as

$$z = 10^{s*B(z)} \qquad \text{(Equ 10)}$$

The constant s is the bandwidth optimisation factor which leads to the evaluation of $B(z)$. $B(z)$ is the effective bandwidth as defined in the theory described below.

Figure 4 shows the price curve and the effective bandwidth curve. At connection set-up the customer declares a value for z, z_d, as part of the traffic descriptor. It is assumed that the customer has a prior knowledge of the source characteristics (source rate distribution, R) which is necessary for calculation of the effective bandwidth used in (Equ 10). The declaration of z_d enables the network operator to deduce the price curve, $Price(z)$, which is drawn as a straight line tangent to the standard effective bandwidth curve $B(z)$ at the point where $z = z_d$. The price "quoted" back to the customer is based on the effective bandwidth, i.e., Price=α*effective_bandwidth*T.

During the time when the connection is in progress, "scanning" of the actual traffic stream allows the network operator to draw up a new source rate distribution curve. This enables recalculation (estimation) of the effective bandwidth, leading to a new measured value for z, z_m, using (Equ 10). If $z_m=z_d$, the price remains at the quoted value. If on the other hand z_m is greater than the declared value, $z_m=z_{mo}$ in Figure 4, then the price is $(a+bz_{mo})$ per unit time, which results in the penalty as shown in the graph. Equally, if $z_m=z_{mu}$, the price remains at $B(z_d)$ per unit time resulting in over payment.

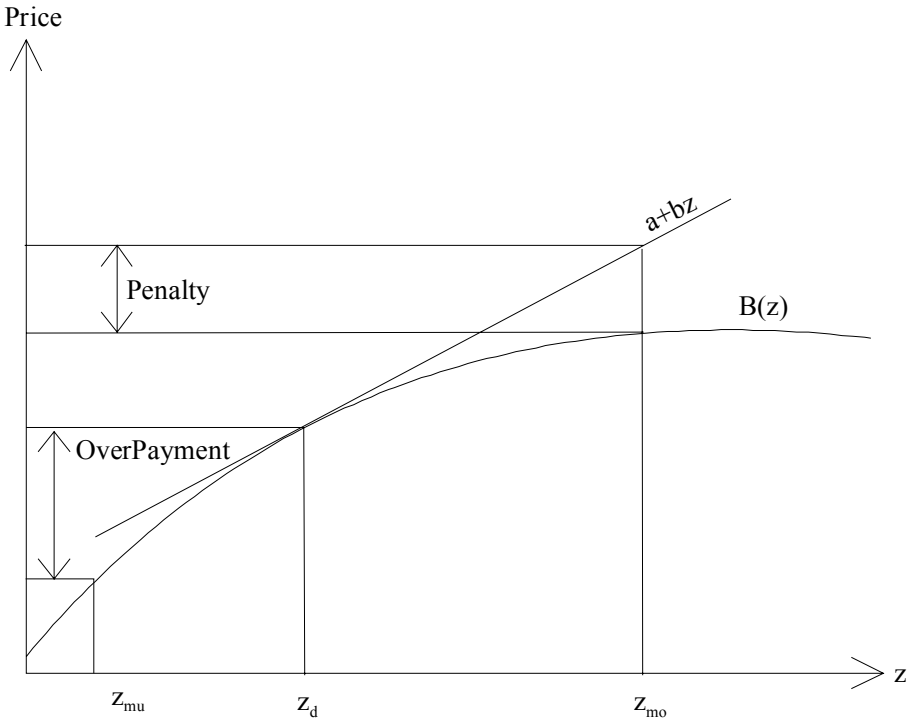

Figure 4: Kelly's price curve.

The Theory of Kelly's Charging Scheme [KELL01]

Suppose that n identical sources share a single resource of capacity C and let $X=x_1,\ldots,x_n$ be the load produced by these sources. The effective bandwidth is then defined as

$$B(z) = \frac{1}{s}\log(z) \qquad \text{(Equ 11)}$$

where s is a constant and z is defined as

$$z = \mathrm{E}e^{sX} = \sum_{i=1}^{C} e^{sx_i} * P(x_i) \qquad \text{(Equ 12)}$$

The parameter s is evaluated as follows. Let

$$M(s) = \frac{1}{s}\log \mathrm{E}e^{sX} \qquad \text{(Equ 13)}$$

It can be shown that for a given set of s and n the condition

$$\max_{n}\left\{\min_{s}\left[nM(s) - sC\right]\right\} \le -q \qquad \text{(Equ 14)}$$

such that

$$\Pr\left\{\sum_{j=1}^{n} x_j > C\right\} \le e^{-q} \qquad \text{(Equ 15)}$$

where q is the cell loss probability parameter (giving CLP=10^{-q}). X_j is the load produced by source j. The value of n and s is the optimisation pair. If

$$nM(s) - sC = -q \qquad \text{(Equ 16)}$$

then,

$$n = \frac{-q + sC}{M(s)} \qquad \text{(Equ 17)}$$

Equating,

$$\frac{\partial n}{\partial s} = 0 \qquad \text{(Equ 18)}$$

for the largest value of n, N, leads to the most optimal value for s. The effective bandwidth is then C/N; C is the switch resource (in bandwidth).

Further detail on the mathematical technique used in the calculation can be found in [COU01][COU02].

3.5.2. Lindberger's Charging Scheme

The general expression for Lindberger's scheme is [LIND01]:

$$Price = \alpha_{\omega,\tau} * d * T \qquad \text{(Equ 19)}$$

where ω defines the geographical location of the destination node (e.g., local, regional, international) and τ defines the time of the day the connection was made (e.g., peak, off-peak, weekend). The parameter $\alpha_{\omega,\tau}$ is the cost (to the customer) per unit bandwidth per unit time for a given ω and τ. d, is the estimated (not declared) effective bandwidth, and T is the duration of the connection. The effective bandwidth is calculated according to the theory proposed in [COST01]. Like Kelly's scheme, the calculation of the effective bandwidth here also relies on scanning of the traffic streams to create the source rate distribution, R.

The Theory of Lindberger's Charging Scheme [LIND01]

The effective bandwidth is defined as

$$d = am + b\frac{\sigma^2}{C} \qquad \text{(Equ 20)}$$

where m is the mean bandwidth and a and b are constants evaluated according to the formulae below. σ^2 is the standard deviation of the source rate distribution, R or *PeakBandwidth(PeakBandwidth-MeanBandwidth)* for the ON/OFF sources. C is the switch resource specified in bandwidth, and the coefficients a and b are defined [COST01] as

$$a = 1 - \frac{\log CLP}{50} \qquad \text{(Equ 21)}$$

$$b = -6\log CLP \qquad \text{(Equ 22)}$$

where the CLP is the cell loss probability.

3.5.3. Botvich's Charging Scheme

Botvich [CC01] takes the notion of effective bandwidth developed by Kelly further to simplify the basis for charging. The argument is that online implementation of the concept (of effective bandwidth) is difficult for network operators to implement, too complex for the customers to understand, and that charges are not predictable at the time of the connection acceptance. The proposal is as follows. The effective bandwidth is evaluated periodically (by the network operator), offline, for each and every connection and is then applied according to the theorem below to derive a constant k. During normal operation, price is based on the expression

$$Price = \alpha * k * m * T \qquad \text{(Equ 23)}$$

where α is the cost (to the customer) per unit bandwidth per unit time. T is the connection duration and m is the declared mean rate. The constant k is defined according to the theory described below.

The Theory of Botvich's Charging Scheme [CC01]

Assuming that during a sufficiently long period of time the network had n connections with mean bit rates $m_1,...,m_n$ respectively. Denote duration of connections to $T_1,...,T_n$ respectively and further assume that the network records these connections and estimates the effective bandwidth $B(z)_1,..., B(z)_n$ periodically. Then

$$k = \frac{\sum_{i=1}^{n}\left(T_i * B(z)_i\right)}{\sum_{i=1}^{n}\left(T_i * m_i\right)} \qquad \text{(Equ 24)}$$

3.5.4. Griffiths' Charging Scheme

With Griffiths' scheme [GRIF01][GRIF02], customers request a particular bandwidth (peak cell rate), called the design rate (*dr*), and the network implements an ATM-filter which ensures that the output of the filter is an estimated equivalent bandwidth (*eb*) which is, at worst, not more than 1.5 times the design rate. If the customer violates the contract and increases the *dr* during the connection without re-negotiation with the network, the filter will discard cells. Customers pay for the *eb*.

The design rate is peak cell rate for CBR sources but no suggestion is made for the design rate for VBR based services. For the purpose of this research, the author makes the assumption that customers will use traffic shapers for VBR sources and declare the shaper leak rate as the design rate.

The filter (shown in Figure 5) is a dual leaky bucket UPC with buffer sizes re-dimensioned and a small pre-buffer introduced prior to the leaky bucket. The cup, saucer and pre-buffer sizes and their leak rates are dimensioned by the network operator both experimentally and using M/D/1/K queuing analysis. This dimensioning work is carried out periodically, offline. The price is based on the equivalent bandwidth through

$$Price = \alpha * eb * T \qquad \text{(Equ 25)}$$

The Theory of Griffiths' Charging Scheme [GRIF01][GRIF02]

Full description of the theory can be found in [GRIF01][GRIF02]. To summarise, the configuration of the filter is as shown in Figure 5.

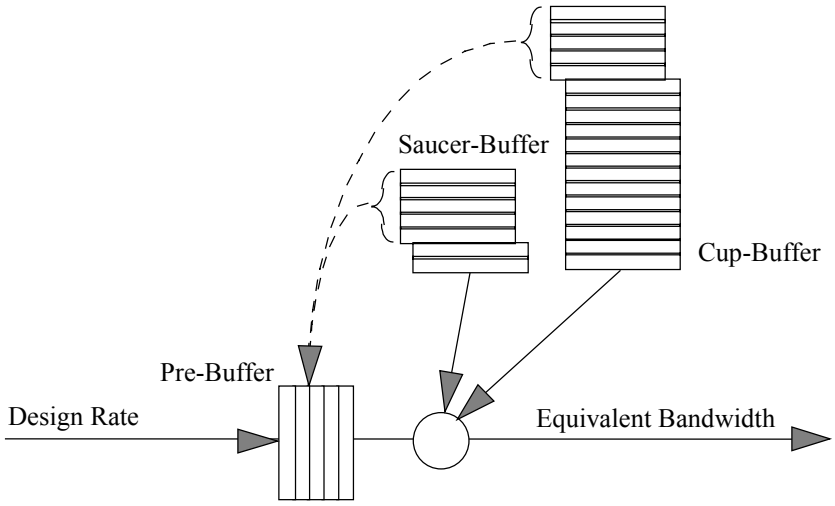

Figure 5: Griffiths' ATM filter.

There are two algorithms: calculation of the cup and saucer leak rate; and determination of cup, saucer and pre-buffer sizes. Procedure for determination of the buffers is given in [GRIF01][GRIF02]. For design rates of 0.001 to 0.1, the cup and saucer sizes are 17 and 1 respectively with pre-buffer size of 6 cells. Design rates falling outside these rates require the dimension to be re-evaluated using the proposed theory.

3.6. Qualitative Evaluation of the Existing Charging Schemes

First and foremost, all the charging schemes discussed so far are applicable only to VBR based services; the respective authors do not offer suggestion for ABR or UBR transfer capabilities. The Griffiths' scheme is also applicable to CBR traffic while the rest assume charges based on peak cell rate for CBR capability.

Two parameters are at the heart of the first scheme (Kelly's): the declared effective bandwidth and the measured effective bandwidth. If the actual measured value is equal to the declared value the price remains at the quoted value. If, on the other hand, the declared value is lower than the measured value, then the price becomes $a+bz$ resulting in significant increase in the actual amount payable from the price quoted at the time of the connection. This will occur in addition to possible cell losses due to the UPC function.

If the declared value is higher than the measured value then the price remains at the quoted value, resulting in overpayment. This is, however, necessary since extra resources will have been allocated by the CAC and must be paid for. Therefore:

- the customer must provide an accurate estimation of the effective bandwidth;

- the measurement of the effective bandwidth must be accurate;

However, due to the statistical nature of the traffic characteristic, the source rate distribution is likely to vary each time it is computed. This could result in variation of the effective bandwidth between the declared and measured value. The result could be that the customer under or overestimates the effective bandwidth leading to severe penalty or overpayment. The implication of this is that the price deduced by the scheme can be hard to predict.

In addition, the complexity involved in computation of the effective bandwidth and the evaluation of the price using the price curve (Figure 4) makes the scheme unclear from the customer point of view. Kelly's scheme also suffers from the issue of control, choice and flexibility criteria. The scheme does not provide a *clear-cut* choice to the customers in terms of price categories for different quality of services. Likewise, the scheme cannot easily be adopted to support all four of the ATM transfer capabilities (CBR, VBR, UBR and ABR): only VBR is accounted for. Furthermore, pricing cannot be used as a control mechanism for network utilisation optimisation and maximising profitability. The penalty mechanism performs more as an UPC function than a network utilisation and profit maximisation tool.

In the second charging scheme, Lindberger's, the tangent curve $a+bz$ is removed and charge is simply based on the measured effective bandwidth, d, alone. This removes the penalty and overpayment element. However, due to the online calculation of the standard deviation value using the source rate distribution, which can vary due to the statistical nature of the traffic characteristics, the effective bandwidth estimated may deviate from that of the customer's expectation. If the estimation is above or below the declared value then charges are made on the estimated bandwidth. If the estimated value is above the declared bandwidth then the UPC, which is set-up based on the declared value, might discard cells and the price payable will be increased. If on the other hand, the estimated value is below the declared bandwidth, extra resources will have been allocated by the CAC and will not be paid for by the customer.

Lindberger's scheme also introduces difficulties similar to that of Kelly's scheme; i.e., clarity, flexibility, control and choice are not an inherent capability of the charging scheme.

In the third, Botvich's scheme, the author of the scheme removes the need for the rate distribution calculation on-line. Price is based on the estimated bandwidth calculated by multiplying the mean rate by the parameter *k*, based on Kelly's effective bandwidth algorithm, which is evaluated offline periodically by the network operator. After every evaluation of *k*, customers can be informed of this in advance to remove any uncertainty. Customers only need to declare the mean rate depending on their requirements and the charge will be known. UPC can be set up to ensure that *k*declared-mean* is conformed to.

With Botvich's scheme, if the actual bandwidth is below the declaration (k*mean) then the UPC might again discard cells, but with this scheme charges will not increase. If the actual bandwidth is above the declaration then the price will remain at the declared value (overpayment) since resources will have been already allocated by the CAC and must be paid for.

Now, the question is how often should *k* be evaluated, when should it be evaluated, and who does the evaluation and how accurate is it? It has already been argued that the effective bandwidth can vary at any time, which will have an impact on the *k* parameter. If the real *k* parameter for a connection differs from the *k* parameter that was calculated offline some time earlier, will this affect the action of the UPC? Customers would also want to know how *k* is evaluated. Finally, a question that could be asked is would a new contract be required every time *k* is re-evaluated?

The fourth (Griffiths') scheme completely does away with the effective bandwidth and the mean rate notion. Customers are expected to request a bandwidth (for CBR and VBR sources) depending on their requirements and willingness to pay for it. The network places a filter to ensure that this bandwidth is not exceeded. The filter rate, called the equivalent bandwidth, closely matches the requested bandwidth and the customer pays for that equivalent bandwidth. The estimated equivalent bandwidth for various requested bandwidths can be stored in a database and, therefore, when a request is made exact chargeable bandwidth can be announced in advance to avoid unpredictability.

If the bandwidth estimation is below that declared, then the filter might discard cells but charges will not be increased. If the estimated bandwidth is above the declaration then the price is based on the estimated value (overpayment) since resources will have been already allocated by the CAC and must be paid for.

The problems that can be identified with Griffiths' scheme (indeed applying equally to Botvich's scheme) are one of control, choice and flexibility. Neither schemes allow control in terms of ability to use pricing to better utilise the network resources and maximise profitability. Likewise, customers cannot *easily* choose different QoS parameters to minimise their network cost. Furthermore, the schemes cannot be adopted to support all four of the ATM transfer capabilities (VBR, CBR, UBR and ABR) – only VBR and CBR capabilities can be supported.

Comparing to the criteria discussed earlier one can observe the following: the first two schemes fail on criteria (1), (3), (4), (6) and (7). The third and fourth schemes fail on criteria (4), (6) and (7).

	Clear	Account.	Predict.	Flexible	Practical	Control	Choice
Kelly	✗	✓	✗	✗	✓	✗	✗
Lindberger	✗	✓	✗	✗	✓	✗	✗
Botvich	✓	✓	✓	✗	✓	✗	✗
Griffiths	✓	✓	✓	✗	✓	✗	✗

Table 5: Qualitative evaluation of charging schemes.

Table 5 summarises the qualitative evaluation of the existing charging schemes. The boxes marked with a cross indicate that the scheme does not fulfil the specified criteria.

The new QoS based charging scheme is designed to meet all the criteria defined in this table.

3.7. SUMMARY

This chapter provided the background information relating to charging for ATM. Specifically the following items were addressed:

- The general ATM charging schemes.
- Importance of QoS and the mechanisms for mapping QoS from customer perspective onto parameters used by the network and the charging schemes.
- The charging scheme evaluation criteria.

Also presented was a description of four of the most prominent existing charging schemes in the literature and an assessment of them based on the criteria developed. In chapter 4., the author develops a new charging scheme designed to meet all the criteria identified and address the shortfalls associated with the existing charging schemes.

4. A NEW APPROACH: QOS BASED CHARGING SCHEME

4.1. OVERVIEW

In the previous chapter (section 3.2.) a general charging scheme was presented (consisting of a set of service, subscription and network related costs), with (Equ 2) defining the variable part of the cost. That is: Price=α*β*T representing the charges for network resource usage and forming the basis for requiring resource usage based charging schemes.

In this chapter, the focus is on resource usage based charging, particularly on the development of a new approach called the QoS based charging scheme.

The previous chapter also provided a description of four existing resource usage based charging schemes and some of the problems associated with them. In this chapter, the author describes the new scheme, which is designed to meet these problems.

4.2. CUSTOMERS VIEW OF THE QOS BASED CHARGING SCHEME

From the customer point of view, *simplicity*, *choice* and *predictability* are the key features of the QoS based charging scheme [MIAH01].

Customers declare, for the VBR based services:

- resource required, β (in terms of $y*m$, where m is the mean cell rate of the traffic source and $y=1..5$)
- QoS (ζ)
  - ♦ cell loss probability (CLP)
  - ♦ maximum cell transfer delay (CTD)
- time of the day at which the connection is to be made (τ)
- geographical location of the destination switch (ω).

For the CBR based services, the declarations are the same except that β is equated to the peak cell rate of the traffic source. Likewise, for the ABR based services, the declarations are the same except that β is based on both the minimum cell rate and the volume of cells transmitted.

For the UBR based services however, β is equated to the volume of cells transmitted and therefore the declarations are:

- time of the day at which the connection is to be made (τ);
- geographical location of the destination switch (ω).

In all cases, given the declaration the network responds with a price. If the customer accepts the offer, this price remains valid until the end of the connection. The principle of the scheme is as follows: a set of QoS streams is made available to the customers, each dimensioned to support particular QoS requirements. The price of usage of each of the QoS streams is dependent on the declaration and is announced as price per unit of resource, β, per unit time. Different QoS stream will have different prices.

4.3. Operators view of the QoS Based Charging Scheme

From the network operator point of view, *control* and *flexibility* are the key features of the QoS based charging scheme [MIAH01], allowing:

- network utilisation and profit maximisation;
- ability to support different transfer capabilities (which in turn enables support of different services).

The architecture of the switch conforming to the scenario assumed for the scheme is shown in Figure 6. (This is a conceptual view of the architecture since the actual hardware does not necessarily have to be in that configuration although Figure 6 does represent the top-level system view of a real ATM switch developed by a well-known switch manufacturer.)

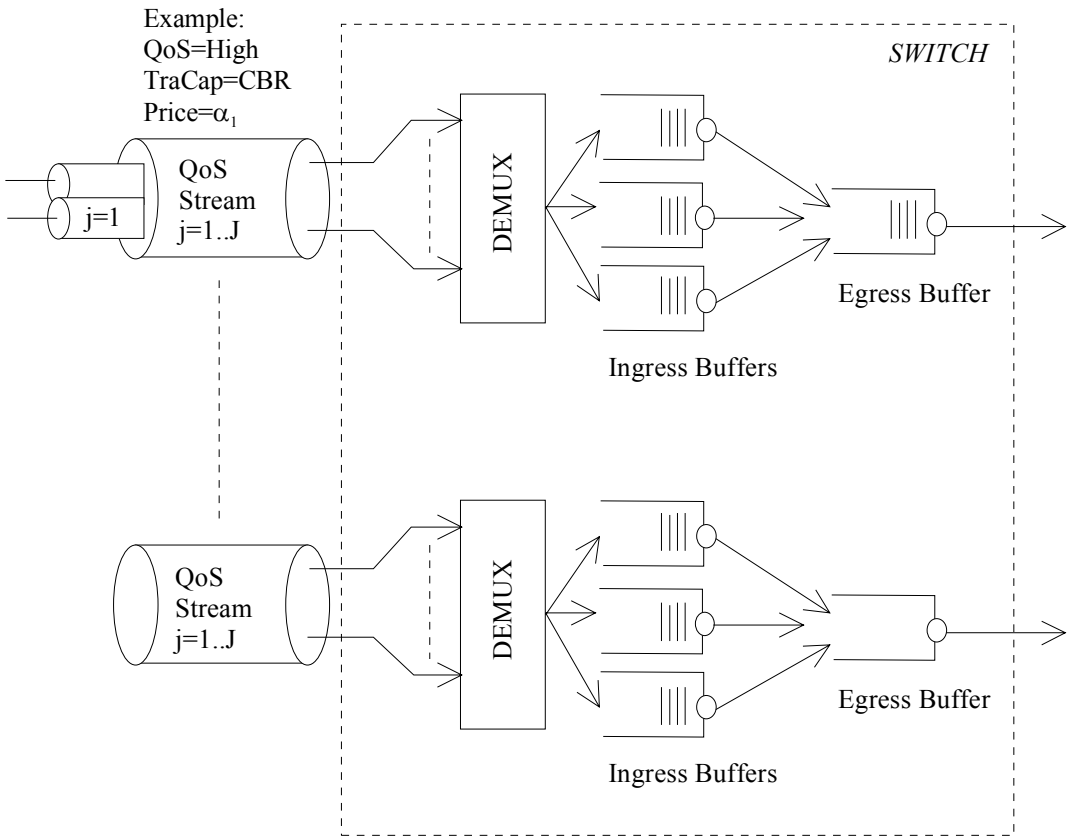

Figure 6: Top-level switch architecture of a real ATM switch.

Each ingress buffer in Figure 6 supports a particular QoS stream that the customer chooses. The streams are dimensioned with a set of buffers (ingress and egress), including a traffic shaper (not shown in Figure 6) positioned immediately before the ingress buffer, for VBR based services. If the shapers are implemented at the customer premises, cheaper CBR capability can be purchased.

Buffers (ingress and egress) are dimensioned using the M/D/1/K and ND/D/1 queueing analysis since:

- it provides a good approximation of buffer behaviour;
- it is simple to calculate and easy to implement;
- it is a tried and tested approach [PITTS01].

Price quoted to the customers for usage of the QoS stream j is deduced from the cost $\alpha_j(\zeta,\beta,\tau,\omega)$ per unit β per unit time. This price will vary according to the customer declaration and the buffer dimensioning (different QoS streams will have different prices).

4.4. DESCRIPTION OF THE QOS BASED CHARGING SCHEME

Consider the scenario depicted in Figure 7. In order to devise a charging scheme it is necessary to go back few steps and look at the activity of the network functions which determine whether or not a connection should be admitted into the network. By identifying all the issues involved in making this decision a charging scheme can be constructed.

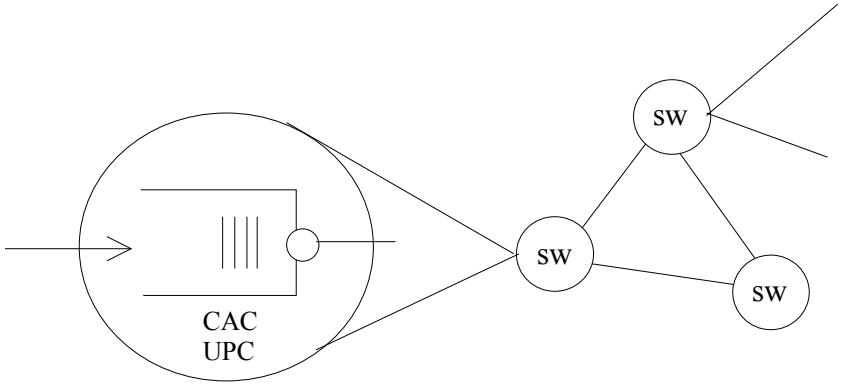

Figure 7: Key ATM network ingress node functions.

The CAC is responsible for determining whether or not a connection can be admitted into the network given the traffic descriptor; the UPC is responsible for ensuring that the connection maintains the declared traffic parameters. Currently, there are no real practical CAC algorithms that can evaluate the resource requirement (bandwidth) accurately prior to the connection being made and so most algorithms simply base their decision on the peak cell rate (peak rate allocation).

The effective bandwidth algorithms used by Kelly and Lindberger can, in many cases, estimate the amount of bandwidth to allocate. These however, like all other effective bandwidth algorithms, operate on the same principle: given the traffic descriptor and the buffer size (see Figure 8) the algorithm evaluates an effective bandwidth that the CAC can use for resource allocation. However, the problems associated with these calculations are that:

- It is difficult to provide accurate information on the source characteristics prior to the connection being admitted into the network. The source rate distribution, R, required as part of the traffic parameter is not always known in advance.
- The buffer considered in the calculation supports multiple connections with varying QoS requirements. The effective bandwidth algorithm must, therefore, take into

account the most stringent QoS requirement for all the connections, resulting in overestimation of bandwidth and price for connections with low QoS requirement.

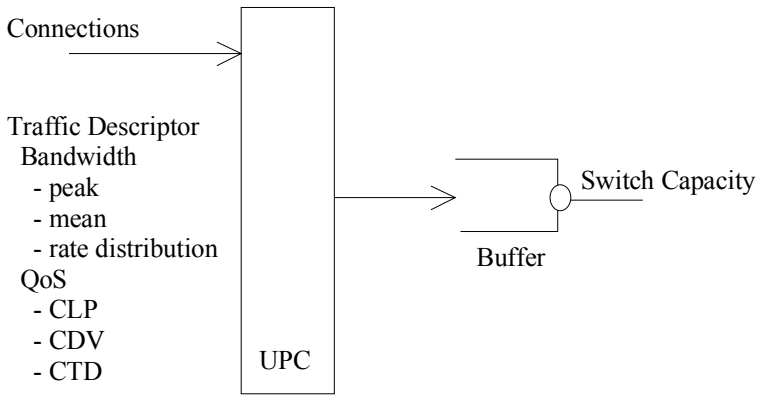

Figure 8: Typical connection set-up process.

In order to simplify the connection set-up process and the charging mechanism one can transform the configuration shown in Figure 8 into a set of network buffers as shown in Figure 9. With the new arrangement, the total bandwidth, C, available from the switch is distributed among a set of QoS stream, j, such that each stream has a net bandwidth $net\_ingress\_C_j$. The distribution of the bandwidth, C, is carried out through the following transformations:

$$egress\_C = C * egress\_\rho \quad \text{(Equ 26)}$$

$$ingress\_C_j = ingress\_bd_j * egress\_C \quad \text{(Equ 27)}$$

$$net\_ingress\_C_j = ingress\_C_j * ingress\_\rho_j \quad \text{(Equ 28)}$$

The definitions of the parameters are as follows:

1. C is the total bandwidth available from the switch.
2. $egress\_C$ is the egress buffer service rate (in bandwidth).
3. $egress\_\rho$ (varying from 0 to 1) is the egress buffer utilisation.
4. $ingress\_C_j$ is the ingress buffer (j) service rate (in bandwidth).
5. $ingress\_bd_j$ (varying from 0 to 1) is the bandwidth distribution factor.
6. $ingress\_\rho_j$ (varying from 0 to 1) is the ingress buffer (j) utilisation.
7. β in Figure 9 is the resource usage information as defined in section 4.2.
8. α_j in Figure 9 is the cost of unit of resource per unit time for QoS stream j.
9. α in Figure 9 is the overall cost, based on market strategy (see section 1.1.2.).

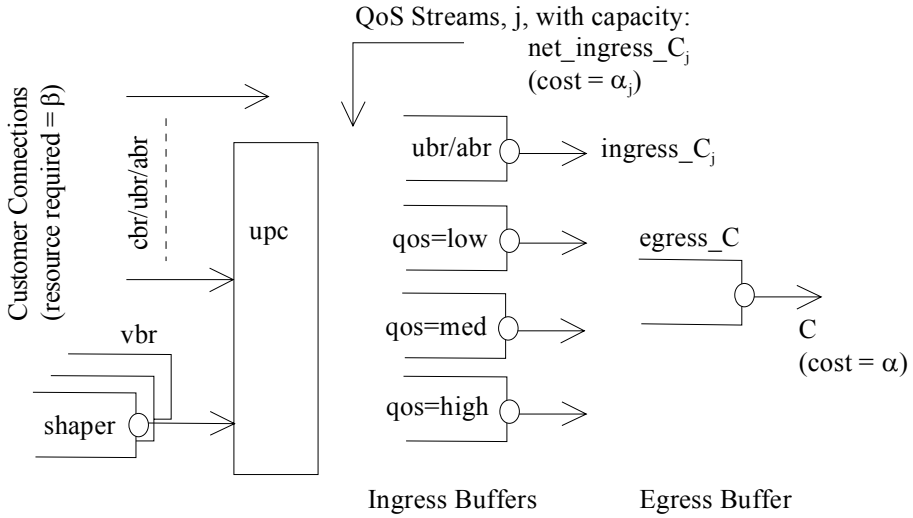

Figure 9: Segregation of traffic streams based on QoS.

egress_ρ determines how much of the total available bandwidth, *C*, that can be made available for usage given the most stringent QoS (CLP) that the switch need to support. If *egress_ρ* is 1.0 then the entire bandwidth, *C*, is made available to the ingress buffers; that is:

$$\sum_{j=1}^{N} ingress\_C_j = egress\_C = C \qquad \text{(Equ 29)}$$

The proportion of *C* that is made available to the stream *j* is also dependent on $ingress\_bd_j$ (Equ 27). This parameter is determined by the management system of the network and is subject to alteration at any time. The objective is to allocate and redistribute the total available resource according to demand and profitability of the QoS streams. For example, during business hours streams with high QoS (which are more expensive) may be allocated the larger part of *C*. This will enable full control of the network to optimise utilisation and maximise profitability.

$ingress\_\rho_j$ determines how much of the bandwidth, $ingress\_C_j$, is actually usable by the stream *j*, given the QoS the stream is supporting (Equ 28). This is the bandwidth efficiency factor (determining how efficient a QoS stream is), and will differ for different streams according to their QoS requirements.

The actual amount of bandwidth available to the QoS stream *j* is $net\_ingress\_C_j$ and is dependent on *egress_ρ*, $ingress\_\rho_j$, $ingress\_bd_j$ and *C*. *egress_ρ* and $ingress\_\rho_j$ are the efficiency parameters and which are also used to deduce α_j from α.

In order to explain the cost allocation and the price deduction for individual connections consider the example where a network switch has:

- an overall bandwidth of *C*;
- cost (α) of access to the network via this switch is £100.00 per unit bandwidth per unit time, which is calculated to meet the target revenue generation determined by business strategies.

If the switch configuration is as shown in Figure 8, then only a single buffer is present in the architecture, which is dimensioned to support the highest QoS a connection could require. This results in the efficiency of the buffer being less than 100% if the number of simultaneous connections is large. For example, to support a CLP of 10^{-8} with the number of connections in the order of 100s, the efficiency may only be 70% with buffer size of 15 cells. Then, in order to meet the target revenue, the cost of the bandwidth usage must be scaled up to cater for the reduction in resource availability of the buffer. The new cost is £142.86 (100/0.70) per unit bandwidth per unit time.

If on the other hand the switch configuration is as shown in Figure 9, then the efficiency of each of the QoS streams is different since connections with different QoS requirements are segregated among the ingress buffers (efficiency of the egress buffer is 100%). Let the efficiency of the four streams shown in Figure 9 be 70, 80, 90 and 100% for high, medium, low and UBR QoS service streams. Then, the costs (α_j) of the bandwidth made available to the customers through each of the streams (ingress buffers) are £142.86, £125.00, £111.11 and £100.00 per unit bandwidth per unit time.

With both the arrangements, the net revenue generated by the network operator remains at the target value. However, the cost of the bandwidth usage to the customers varies between the two approaches. In the first case all customers pay £142.86. *In the second case, customers have a choice in selecting lower QoS to reduce the cost of their network usage. In a competitive environment, this is significant.*

With the QoS based charging scheme, price of connections using stream *j* is

$$Price = \alpha_j(\zeta, \beta, \tau, \omega) * \beta * T \qquad \text{(Equ 30)}$$

for both the CBR and VBR (real and non-real-time) connections and for UBR connections

$$Price = \alpha_j(\beta, \tau, \omega) * \beta + ct \quad \text{(Equ 31)}$$

The ABR connections can be charged using the CBR expression up to the minimum cell rate; anything above that is based on UBR charging minus the *ct* part.

The function $\alpha_j(\zeta,\beta,\tau,\omega)$ is the cost per unit of bandwidth per unit time for usage of stream *j* and is a function of the QoS requirement (ζ), resource requirement (β), and the time of the day at which the connection was set-up (τ) as well as the geographical location of the destination switch (ω). *T* is the duration of the connection and the element *ct* represents a small incremental price to discourage customers from setting up a UBR connection and keeping it open indefinitely without transmitting any data (since without transmission of any data the price will be null).

The resource requirement parameter, β, is specified in bandwidth for CBR/VBR/ABR traffic and volume of cells transmitted for UBR traffic.

4.5. Theory of the QoS Based Charging Scheme

The QoS based charging scheme involves the following calculations:

1. dimensioning of the egress buffer;
2. dimensioning of the ingress buffers;
3. configurations of the QoS streams;
4. determination of the cost function;
5. evaluation of β.

The calculations are bounded by the following constraints:

1. the overall CLP values of all the connections remain within the required limit;
2. the overall CTD values of all the connections remain within the required limit;
3. the egress buffer must be transparent in terms of efficiency and delay (i.e., *egress_*ρ*=1.0* and the delay introduced by the egress buffer must be negligible).

4.5.1. Egress Buffer Dimensioning

The egress buffer has a service rate (bandwidth) of *C* and supports *N* number of connections; each connection is defined as a QoS stream (output of the ingress buffers), as shown in Figure 9. In order to calculate the dimension of the buffer, let

$$egress\_\rho = 1.0 \qquad \text{(Equ 32)}$$

i.e., no bandwidth is wasted due to low efficiency (utilisation) at the egress buffer (constraint 3). Then (from Equ 29 and Equ 26 to Equ 28),

$$C = \sum_{j=1}^{N} ingress\_C_j \qquad \text{(Equ 33)}$$

The objective now is to deduce the CLP (cell loss probability) that would be introduced by the buffer for different values of *N* and buffer size *egress_b*.

Assuming that the maximum number of QoS streams, *N*, is small, the ND/D/1 queueing model is appropriate. According to [PITTS01], such a queueing mechanism has the property that the probability of the queue size exceeding *egress_b* (and therefore the CLP) is

$$Q(egress\_b) = \sum_{n=1}^{N} \left\{ \frac{N!}{n!(N-n)!} \left(\frac{n - egress\_b}{D} \right)^n \left[1 - \left(\frac{n - egress\_b}{D} \right) \right]^{N-n} \frac{D - N + egress\_b}{D - n + egress\_b} \right\} \qquad \text{(Equ 34)}$$

This leads to the result shown in Table 6. The first column shows the size of the buffer whilst the first row shows the CLP values. For example, if the highest QoS a connection is likely to require is CLP of 10^{-12} and the number of QoS streams required (*N*) is 10, then, *egress_b=10* will satisfy *egress_ρ=1.0*.

	CLP											
b	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-11}	10^{-12}
5	23	11	8	6	5	5	5	5	5	5	5	5
10	89	45	30	23	19	16	14	13	12	11	11	10
15	200	100	67	50	41	34	30	26	24	22	20	19
20	353	176	118	89	71	60	52	45	41	37	34	32
25	550	275	183	138	111	92	80	70	63	57	52	48
30	790	395	264	198	159	133	114	100	89	81	74	68
35	1064	537	358	269	215	180	155	136	121	109	100	92
40	1389	701	467	351	281	234	201	176	157	142	129	119
45	1758	886	591	443	355	296	254	223	198	179	163	150
50	2171	998	729	547	438	365	313	275	244	220	201	185

Table 6: Number of connections for a given CLP and buffer size (b) for ND/D/1 queue.

In terms of the other QoS parameter, CTD, the maximum delay that is introduced by the egress buffer is

$$egress\_ctd = 424 * \frac{egress\_b}{C} \quad \text{(Equ 35)}$$

where 424 is the length of the ATM cells in bits and *C* is in bits/s.

Simulations performed to test this theory proved to closely match the analytical values presented in Table 6. Full description of the simulation tools used and how they were validated is provided in chapter 5. Table 7 shows the simulation results obtained by multiplexing different number of CBR sources through buffer of varying cell sizes, at loading of 100%.

	CLP											
b	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}	10^{-11}	10^{-12}
5	30	15	10	8	8	7	6	5	5	5	5	5
10	93	51	33	25	23	18	16	12	11	11	10	10
15	219	115	72	57	46	37	33	28	24	22	20	15
20	359	179	123	94	77	63	55	48	43	38	35	30
25	557	277	185	139	115	95	83	72	65	58	52	50

Table 7: Number of connections for a given CLP and buffer size (b) in simulations.

4.5.2. Ingress Buffers Dimensioning

There are *N* ingress buffers for *N* QoS streams. Each buffer has a service rate (bandwidth) of $ingress\_C_j$ (Equ 27). The objective is to determine the percentage of this rate that is available for use given the QoS that the stream is supporting, i.e., the efficiency of the buffers ($ingress\_\rho_j$).

Assuming that the number of connections entering an ingress buffer, *j*, is large enough to equate them to a single stream of Poisson traffic with arrival rate λ, the M/D/1/K queueing analysis is appropriate. Let the lost traffic on stream *j* be

$$lt = rx - tx = \mathrm{E}[a] - ingress\_\rho_j = \mathrm{E}[a] - \{1 - s(0)\} \quad \text{(Equ 36)}$$

where *rx* is the offered traffic and *tx* is the carried traffic. *E[a]* is the average arrival rate and *s(0)* is the probability of zero cell in the buffer. According to [PITTS01]

$$s(0) = \frac{1}{\sum_{k=0}^{ingress\_b_j} u(k)} \quad \text{(Equ 37)}$$

where *ingress_b*$_j$ is the size of the ingress buffer *j*. The CLP achievable from this buffer is then

$$CLP = \frac{E[a] - \{1 - s(0)\}}{E[a]} \quad \text{(Equ 38)}$$

with

$$ingress\_\rho_j = \frac{\lambda}{igress\_bd_j * C} \quad \text{(Equ 39)}$$

To deduce *u(k)*, M/D/1/K state probability is used. [PITTS01] showed that with a M/D/1 queue of infinite buffer size the probability of *k* cells in the buffer is

$$s(k) = \frac{s(k-1) - s(0)a(k-1) - \sum_{i=1}^{k-1} s(i)a(k-i)}{a(0)} \quad \text{(Equ 40)}$$

with

$$a(k) = \frac{\lambda^k}{k!} e^{-\lambda t} \quad \text{(Equ 41)}$$

where *t* is a fixed duration of time, equated to 1 for a single cell slot in discrete mode, and

$$s(0) = 1 - E[a] \quad \text{(Equ 42)}$$

However, for a finite buffer size, the zero state probability (i.e., probability of no cells in the buffer), *s(0),* cannot be determined through (Equ 42) since now there may be losses at the buffer due to finite buffer size. Therefore define

$$u(k) = \frac{s(k)}{s(0)} \quad \text{(Equ 43)}$$

i.e., *s(0)=1*. This gives

$$u(k) = \frac{u(k-1) - s(0)a(k-1) - \sum_{i=1}^{k-1} u(i)a(k-i)}{a(0)} \quad \text{(Equ 44)}$$

$$u(k) = A(X) \quad \text{(Equ 45)}$$

$$A(X) = 1 - [a(0) + a(1) + ... + a(X-1)] \quad \text{(Equ 46)}$$

where *A(X)* is the probability of *X* number of cells arriving into the buffer, *X* is the buffer size.

Figure 10 shows the load, *ingress_*ρ_j, allowed by the buffer given the CLP requirement. Two curves are shown on the graph representing buffer sizes of 10 and 15 as an example.

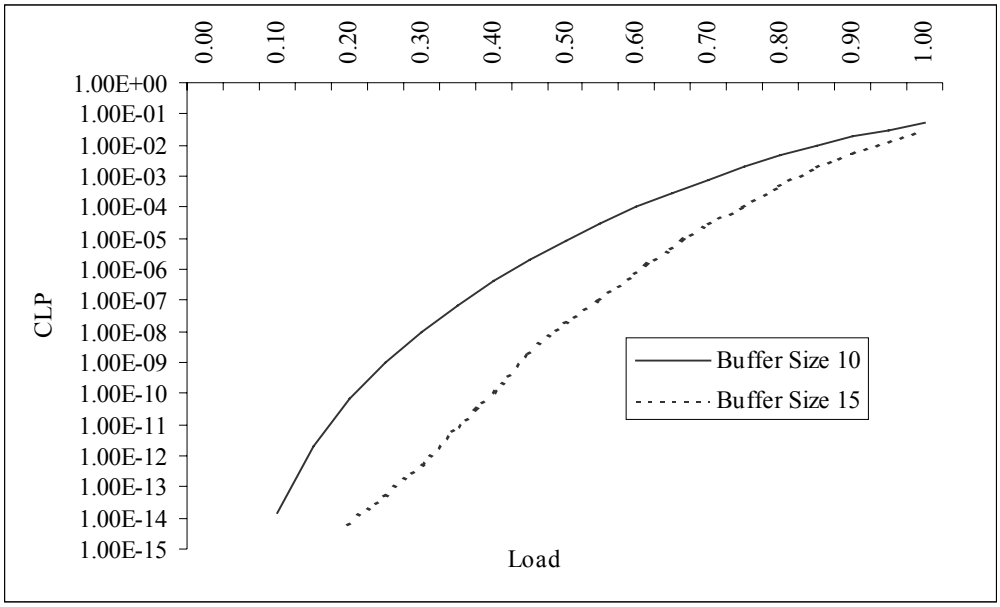

Figure 10: CLP in M/D/1/K queue given the Load and the buffer size.

In terms of the other QoS parameter (CTD), the maximum delay that is introduced by the ingress buffer, *j*, is

$$ingress\_ctd_j = 424 * \frac{ingress\_b_j}{ingress\_C_j} \qquad \text{(Equ 47)}$$

where 424 is the length of the ATM cells in bits and *C* is in bits/s.

Simulations performed to test this theory proved to closely match the analytical values presented Figure 10. Full description of the simulation tools used and how they were validated is provided in chapter 5. Figure 11 shows the simulation results obtained by multiplexing a Poisson traffic stream through buffer of cell sizes 10 and 15.

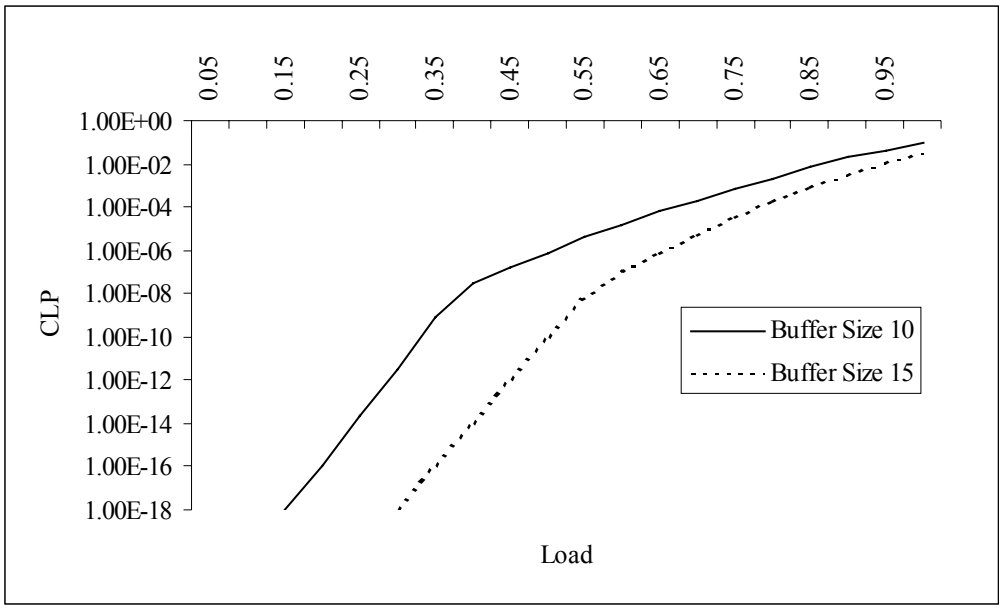

Figure 11: CLP in a single simulated queue given the Load and the buffer size.

4.5.3. Configuration of QoS Streams

Configuring the QoS streams is an iterative process whereby buffers are dimensioned with some assumptions (described below) and then the overall QoS of the streams are checked for conformance. If it does not conform then the process is revisited again to a different set of assumptions.

To illustrate this process suppose that 10 different paths are necessary to support:

1. three streams for high, medium and low QoS using real-time VBR and CBR transfer capability;
2. three streams for high, medium and low QoS using non-real-time VBR and CBR transfer capability;
3. two QoS streams using ABR transfer capability;
4. one QoS stream using UBR transfer capability;
5. one QoS stream reserved for other uses.

Suppose also that high medium and low QoS represents CLP values of 10^{-8}, 10^{-6} and 10^{-4} respectively.

The above assumptions are provided only as an example. However, it can also be considered as a general case for switch configuration.

In order to configure these 10 QoS streams the first step is to set-up the egress buffer. Using (Equ 34) and (Equ 35) with *egress_b=10* and *C=100Mbit/s*, the following values are achievable:

1. CLP=10^{-12} at the egress buffer, which is enough to support the most stringent CLP requirement.
2. Delay introduced by the egress buffer, *egress_ctd*, 43us.
3. *egress_*ρ*=1.0.*

Moving onto the ingress buffers, let the buffer sizes, *ingress_*b_j, of each of the streams be:

1. 10 for ABR streams;
2. 10 for real-time streams;
3. 15 for non-real-time streams;
4. 30 for UBR stream;
5. 60 for reserved stream.

Given the required values of CLP for high, medium and low QoS, *ingress_*ρ_j and *ingress_*ctd_j is evaluated for each and every stream using (Equ 38) and (Equ 47). A check is then performed to determine if *egress_ctd*+*ingress_*ctd_j is less than or equal to QoS parameter CTD (taking into account that shapers may also be necessary for VBR sources).

If this is not the case, then a reduced size for the ingress buffers must be assumed and the calculation is performed again. If the condition is meet, the parameter *ingress_*ρ_j is then used to determine the cost function α_j associated with each of the stream.

Notice that the size of the ingress buffers is rather small. This is possible due to segregation of the traffic among different QoS streams. In the above example, the combined buffer size of the 10 streams is 200 cells, which is comparable to a single buffer configuration. With the reduced ingress buffer size, provision is made for using the excess buffer size to introduce traffic shapers, which is necessary for VBR based services. (Note also that a switch may contain many combinations of QoS streams for large number of different customers and services, resulting in buffer spaces running into 1000s.)

The above example was set-up in a simulation tool with tagged input CBR and VBR sources and background load. The simulation was set to run for a prolonged period (10^{-21} time slots). The objective was to see if the tagged sources receive QoS (CTD and CLP) that

is guaranteed by the QoS streams the sources are using. *The result was positive; the CTD and CLP values were slightly better than that predicted by the analytical calculations.* (Improvement of 5% in the efficiency of the QoS streams for the target CLP values was noticed. Only 25% of the cells experienced the maximum delay determined by the theory, the rest had much lower delays.)

Note that the queueing analyses that have been used are for cell-scale queueing, while the ABR and the UBR capabilities may cause burst-scale queueing. This, however, is not a problem, since the queueing analysis is just a tool that is used. If the QoS based charging scheme must, in reality, be used with bursty data traffic, then the user of the scheme can simply substitute the appropriate queueing formula (such as those in [SCH01][PITTS02]) in place of the ND/D/1 or M/D/1/K approach.

4.5.4. The Cost Function

The cost function α_j, for QoS stream j, is a function of the quality of service parameters (ζ), the resource requirement (β), the time of the day a connection is made (τ), and the geographical location of the destination switch (ω); see Equ 30.

Location Dependency

The dependency on ω could be based on the number of switches a connection traverses. This however is unrealistic and a more acceptable approach would be to divide ω into a set of geographical locations, e.g., "local", "regional", "international-1" and "international-2", if a distance component is required.

Note however, that there is an increasingly widespread view that distance related charging is not appropriate; costs should instead be related to density of traffic along different routes rather than distance involved.

QoS Dependency

The ζ parameter determines which ingress buffer (QoS stream) a connection uses. The principle is as follows. Define the total cost of the overall bandwidth, C, as α per unit time (based on the cost of the resource provisioning and business strategy). The cost of the egress buffer per unit time is then

$$egress\_\alpha = \frac{\alpha}{egress\_\rho} \quad \text{(Equ 48)}$$

Assuming that $egress\_\rho=1.0$ then $egress\_\alpha=\alpha$. The percentage of this cost distributed to the QoS stream, j, is[1]

$$ingress\_\alpha_j'' = ingress\_bd_j * egress\_\alpha = \alpha \quad \text{(Equ 49)}$$

The cost of usage of the ingress buffer for stream, j, per unit time, given the efficiency of the QoS stream, is

$$ingress\_\alpha_j' = \frac{ingress\_\alpha_j''}{ingress\_\rho_j} \quad \text{(Equ 50)}$$

Taking the bandwidth into account as one of the charging unit, the cost of use of the QoS stream, j, per unit bandwidth per unit time is

$$\alpha_j = \frac{ingress\_\alpha_j'}{C} \quad \text{(Equ 51)}$$

This is the cost of usage of each of the QoS stream, j, to the customers. Since $ingress\_\rho_j$ is dependent on:

- the cell loss probability parameter (CLP)
- the ingress buffer size, which is bounded by the maximum cell transfer delay (CTD)

the cost is dependent on the QoS parameters (ζ).

Time Dependency

α_j can also vary to take into account the loss of revenue incurred during a busy hour due to loss of customers as a result of connection blocking. The approach is to use the price as the congestion prevention mechanism. It is proposed that the Erlang's probability of lost calls be used to determine different price bands as shown in Figure 12.

[1] The " and ' notations in the equations are used to distinguish the cost parameter computed by considering different dimensioning parameters in the calculation – see glossary for further explanations.

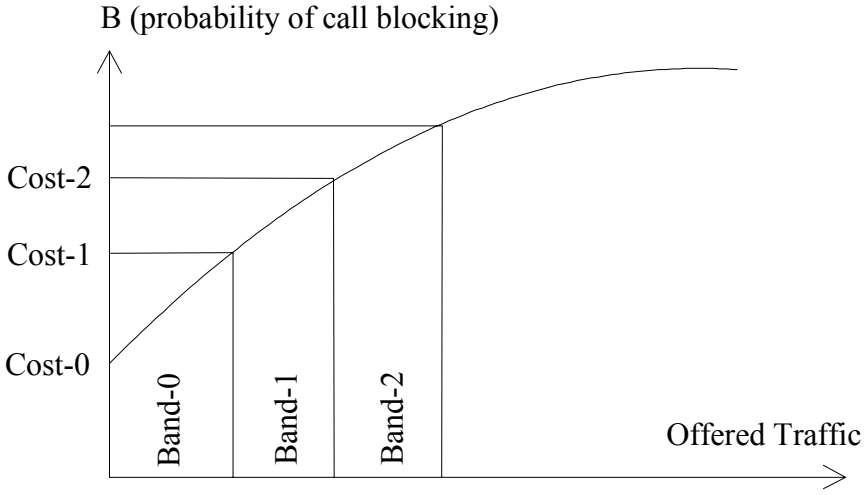

Figure 12: Congestion related price band curve.

The Erlang's formula for call blocking probability is

$$B = \frac{\left[A^N / N!\right]}{\sum_{i=0}^{N}\left[A^i / i!\right]} \qquad \text{(Equ 52)}$$

where N is the number of circuits in the network and B is the probability of call blocking and A is the offered traffic.

A call is equated to a single connection and a circuit is equated to a fixed amount of bandwidth. For example 10Mbit/s per circuit implies $N=15$ for a 150Mbit/s link. *This is a realistic model as proven by extensive research performed with real network and live traffic, in Bell Laboratory, Lucent Technology [BEL01].*

Wholesale Resource Purchase Dependency

The cost of a connection may also vary according to the size of β requested. For wholesale resource purchase customers will expect a reduction in cost. As an example, for a 3 minutes long connection, the price may be $\alpha_j(\zeta,\beta,\tau,\omega)*\beta*3$ producing the graph on Figure 13 (solid line).

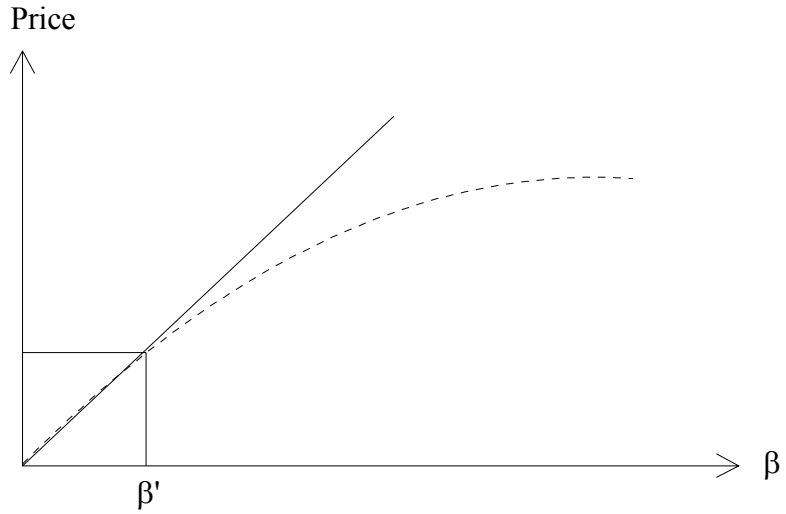

Figure 13: Wholesale resource pricing.

The network can alter the cost such that if β goes beyond a certain value (β'), the cost curve is no longer linear, as shown with the dotted line. The question is, how to determine β'? It is proposed that β' be set to the expected demand on the QoS stream under question, which may be deduced by carrying out analysis such as those shown in [STO01].

For example, suppose that at the time a connection is admitted into the network, the network management system is aware of the expected take up of the available bandwidth on a QoS stream (say 30Mbit/s). Suppose also that the actual amount already taken up, excluding this connection, is 28Mbit/s. If the new connection requests for 5Mbit/s of bandwidth, then the actual price is between $(\alpha_j(\zeta,\beta,\tau,\omega)*2*T)$ and $(\alpha_j(\zeta,\beta,\tau,\omega)*5*T)$. By how much should the curve dip? One approach would be to have decrement of ϕ% for every φ% of bandwidth request above the anticipated demand.

4.5.5. Resource Usage Information

For the VBR based services, calculation of the charging parameter, β, representing the resource requirement, introduces the greatest difficulty. In the existing charging schemes, Kelly, Lindberger and Botvich use effective bandwidth as the basis for resource requirement calculation. Griffiths on the other hand uses peak cell rate.

The difficulties with the effective bandwidth algorithms that were identified are as follows:

1. accurate traffic descriptor information (traffic characteristics such as the source rate distribution) is not available prior to the connection being admitted into the network;
2. use of a common buffer for all the connections requires that the most stringent QoS parameter is taken into account during the bandwidth calculation.

Suppose that the above constraints are somehow eliminated. Could the algorithms then be used to deduce β? To answer this question extensive simulations were performed using the scenario shown in Figure 8 with: buffer size of 300 cells, service rate of 37Mbit/s and target CLP of 1.5×10^{-4}. Three different real-time VBR and CBR sources of the following nature were used in the simulations (full descriptions of the simulation tools and the traffic sources are provided in chapter 5.):

1. ISABEL (see section 5.1.) source for multimedia services;
2. MPEG-I source for video on demand services;
3. CBR source for data/voice (LAN interconnect) services.

The bandwidth evaluated for these sources were:

1. 3.67Mbit/s (ISABEL), 0.85Mbit/s (MPEG) and 2.048Mbit/s (CBR) using the Kelly's effective bandwidth algorithm (Equ 11);
2. 1.91Mbit/s (ISABEL), 0.65Mbit/s (MPEG) and 2.048Mbit/s (CBR) using the Lindberger's effective bandwidth algorithm (Equ 20).

However, the actual experiments showed a number of difficulties in implementing the connections to support these sources. Observe the values given in Table 8 as evaluated by the simulations (the shaded rows show settings that do not conform to the target CLP of 1.5×10^{-4}).

ISABEL	CBR	MPEG	CLP
0	0	43	1.78E-04
0	0	42	1.37E-04
0	1	39	1.02E-04
0	2	36	1.13E-04
0	16	2	1.03E-04
0	17	1	1.13E-04
0	18	0	1.32E-04
1	0	0	0.00E-04
1	0	1	3.83E-04
1	1	0	5.78E-04
2	0	0	7.73E-04

Table 8: CAC bound without traffic shapers for ISABEL, CBR and MPEG sources.

The columns in the table show the number of sources from each of the three categories that could be multiplexed into the buffer given the CLP requirement. The result is summarised below:

1. Maximum of 42 MPEG sources but no ISABEL and no CBR sources could be multiplexed together giving an effective bandwidth of 0.88Mbit/s (37/42) for the MPEG source.
2. Maximum of 18 CBR sources but no ISABEL and no MPEG sources could be multiplexed together giving an effective bandwidth of 2.048Mbit/s (37/18) for the CBR source.
3. Combination of MPEG and CBR sources could be multiplexed together as shown in Table 8.
4. Maximum of 1 ISABEL source and zero MPEG and zero CBR sources could be multiplexed together. The multiplexing problem here could not be improved by increasing the buffer service rate to even 75Mbit/s. On the other hand the service rate could be lowered down to 2.8Mbit/s without degrading the CLP requirement. Hence the effective bandwidth of the ISABEL source could be said to be 2.8Mbit/s.
5. No other sources (MPEG or CBR) could be introduced into the buffer while a single ISABEL source was active.

From the above observations, it can be seen that although the effective bandwidth value calculated by the algorithms closely matches the bandwidth evaluated experimentally (from

Table 8), there are problems associated with realising the actual connections using these values due to the bursty nature of the traffic sources. In the above case, even though the ISABEL source was found to have an effective bandwidth of 1.91-3.67Mbit/s (2.8Mbit/s through simulations), no more than a single source could be multiplexed into a buffer of service rate 37Mbit/s without severely degrading the performance of all other connections sharing the same buffer. *This conclusion was confirmed by subjective tests performed on real network with live traffic (at Ascom, Basel) as part of the European ACTS project EXPERT [EXP01].*

Further complications were observed with the MPEG source through experiment with the real network and live traffic. Although the effective bandwidth estimation by the algorithms and through simulations are generally in agreement with the values between 0.65-0.88Mbit/s, in practice a reasonable quality could only be achieved with a minimum setting of 2.5Mbit/s bandwidth. That is, subjective tests using real network and live traffic showed that a reasonable picture and sound quality could only be achieved with a minimum of 2.5Mbit/s bandwidth per source. *This test was performed by the author using the ORACLE video server running MPEG-I video streams through the Fujitsu APON-R0.1 ATM access network (at Fujitsu Telecommunications Europe Limited, Birmingham). The test set-up configuration is shown in Figure 14.*

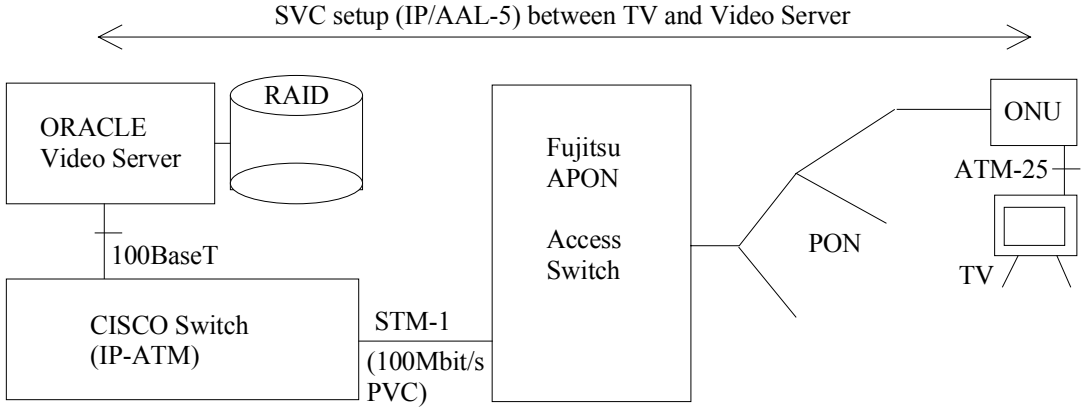

Figure 14: Subjective test set-up for MPEG-I traffic using Access PON.

A scheme such as Kelly's is a "mathematical" effective bandwidth scheme that does not take into account the effect the application has on the resource requirement; i.e., bandwidth needed can be application dependent, rather than just a mathematical manipulation of the traffic profile. The QoS scheme is based on establishing bandwidth to achieve acceptable

QoS and in some cases this must be done by subjective means. Further work in this subjective area is outside the scope of this thesis but is noted in the section on future work.

Given these shortfalls, it can be concluded that for charging purposes at least, where predictability is important, effective bandwidth is not suitable for evaluation of the resource usage requirement. *For the QoS charging scheme, the author proposes that traffic shapers be dimensioned for specific QoS streams that supports VBR traffic.*

To show the benefit of introducing traffic shapers, the buffer size used in the simulation (Figure 8) was reduced to allow additional buffering for the shapers; 10 cells for multiplexing buffer and 290 for traffic shapers, giving a total of 300 as before. The shaper leak rates were set to 2.8Mbit/s (ISABEL) and 2.5Mbit/s (MPEG) while the buffer service rate remained at 37Mbit/s. This scenario was simulated and the result is summarised in Table 9 (the shaded rows show settings that do not conform to the target CLP of 1.5×10^{-4}).

ISABEL	CBR	MPEG	CLP
10	0	5	1.83E-04
10	0	4	1.24E-04
10	1	2	1.08E-04
10	1	3	2.15E-04
11	0	0	1.29E-04
12	0	0	5.12E-02

Table 9: CAC bound with traffic shapers for ISABEL, CBR and MPEG sources.

It can be seen that now 11 ISABEL sources can be multiplexed in to the buffer as well as numerous combinations of ISABEL, MPEG and CBR sources. This was achieved while the maximum cell delay (due to buffering) and CLP remained equal to the earlier case.

The impact of the shapers on the traffic is not in any way different from the buffering considerations that the effective bandwidth algorithms take into account. The effective bandwidth algorithms require a multiplexing buffer, or will work on the basis that cells are lost if more than one arrives at any one-time slot. With the shaper, the buffering is removed from the multiplexer and placed at an earlier stage before multiplexing is carried out.

The final proof of the importance of shapers has been the outcome of the series of experiments carried out by the European ACTS project EXPERT. Setting-up live

connections of voice and video sources over real networks, they observed that far from degrading QoS, traffic shapers can improve statistical multiplexing and network utilisation significantly without loss of perceived QoS [EXP01][EXP02].

Dimensioning of the Traffic Shapers

Given the total sizes of the ingress buffer and the egress buffer of a QoS stream *j*, the traffic shaper for a VBR connection *i* using stream *j* is dimensioned as follows.

$$shaper\_b_{ji} = md_{ji} * \frac{slr_{ji}}{s_{ji} - 1} \qquad \text{(Equ 53)}$$

where s_{ji} is the number of switching stages the connection traverses through, slr_{ji} is the shaper leak rate (in cells per seconds) and $shaper\_b_{ji}$ is the shaper buffer size. This is the formula for generalised process sharing (GPS) applying to fluid "cut through" switching at each stage [COST01]. The parameter md_{ji} is the maximum delay allowed for a connection; i.e., the CTD_i plus the transmission delay. The inclusion of transmission delay allows the model to be applied to a more realistic slope and forward connection switching.

If the maximum delay allowed by the ingress and egress buffers is *ctd_egress* and $ctd\_ingress_j$, given by (Equ 47) and (Equ 35), then the total delay introduced by the ingress switch for stream *j* is[1]

$$ctd\_sw''_j = ctd\_ingress_j + ctd\_egress \qquad \text{(Equ 54)}$$

Any further delay allowed to connection *i* given the maximum cell transfer delay requirement, CTD_i, is

$$ctd\_sw'_{ji} = CTD_i - ctd\_sw''_j \qquad \text{(Equ 55)}$$

Including the transmission delay in the physical medium for connection *i*, $ctd\_transmission_i$, the further delay allowed to the connection within the switch is

$$ctd\_sw_{ji} = ctd\_sw'_j - ctd\_transmission_i \qquad \text{(Equ 56)}$$

Let this equal to the parameter md_{ji} used in (Equ 53) such that

[1] The " and ' notations in the equations are used to distinguish the delay parameter computed by considering different dimensioning parameters in the calculation – see glossary for further explanations.

$$shaper\_b_{ji} = \frac{ctd\_sw_{ji} * slr_{ji}}{s_{ji} - 1} \quad \text{(Equ 57)}$$

with the constraint

$$m_i < slr_{ji} \leq y * m_i \quad \text{(Equ 58)}$$

where y holds values between 1 and 5 and m_i is the mean rate of the source i supporting the connection.

4.5.6. Remarks on CDV

The CDV part of QoS parameters has not been considered in any of the calculation. This has been quite deliberate, as accounting for the definition of CDV is very complex. However, CDV can be considered negligible with respect to buffer dimensioning based on M/D/1/K and ND/D/1 queue. In [COST01] the following conclusions were reached which sum up the impact of the CDV on the QoS streams.

> "If the CDV is negligible, with respect to a Poisson reference process at the network ingress and multiplexing is performed with FIFO queue, subject to the condition that the sum of PCR values is less than the multiplexers rate, then CDV remains negligible throughout the network. This conjecture greatly facilitates the CAC mechanism. The normal load of a multiplexer C and buffer B can be set at the maximum load of a M/D/1/K queue. This can then be made comparable with a sufficiently small probability of a queue length exceeding the buffer size B. connections with negligible CDV are accepted as long as this nominal load is not exceeded on any network link. It remains to specify how long the network can ensure that a connection has negligible CDV at the ingress. Certainly, the simple specification of a CDV tolerance does not characterise the connection sufficiently. The generic connection rate algorithm [proposed by ITU-T] is therefore not a satisfactory policing mechanism in this context. One sure way (perhaps the only way) of ensuring negligible CDV is to actively space the cells of the connection [use of traffic shaper]."

The author, to explore its functionality and the impact it has on the network, has also carried out research into CDV at the ATM layer, the results of which can be found in

[MIAH04]. Briefly, it was shown that the actual CDV (described in terms of standard deviation) introduced by a real ATM passive optical network (Telecom Italy, Iteltel, Milan) is small and that the most significant parameter is the actual delay introduced by the network. Buffering (such as shapers) can be used to remove CDV by ensuring that all cells conform to a maximum delay bounded by the maximum cell transfer delay (CTD).

4.6. SUMMARY

In this chapter, the author described the new QoS charging scheme in detail, with which:

- The total cost of the resource is determined and a cost function for this resource is calculated based on:
  - the actual cost of the resource provisioning;
  - the target revenue generation required by the business strategy.
- The resource and cost function is distributed among a set of QoS streams, based on:
  - the efficiency of the streams (which in turn depends on the QoS supported by the streams);
  - the expected demand on the resources at different QoS provisioning.

It was shown how customers actually buy QoS once the QoS streams have been configured; theories were presented to carry out this configuration. Furthermore, the author describes how the actual price is dependent on the resource requirement, how it is evaluated and how the cost function is dependent on:

- the QoS required by a connection;
- the time of the day a connection is made;
- the geographical locality of the destination switch;
- the size of the resource requested.

Numerous simulations of the multiplexers, shapers and the switch configurations were carried out to validate the theory of the QoS based charging scheme (as mentioned after description of each part of the theory presented within the chapter). In the chapter that follows, the author presents in detail the description of the scenarios simulated and the results obtained for them in order to test, evaluate and compare all the charging schemes.

5. SIMULATIONS OF CHARGING SCHEMES

Four existing charging schemes have been looked at thus far and a qualitative evaluation of them all have been carried out leading to the design of the new QoS based charging scheme. This chapter describes the way in which all the schemes have been simulated and gives the results of these simulations. The objective is to determine the performance of each of the schemes under a typical network scenario. Here, the word "performance" is used to imply how well a scheme quantifies resource usage and meets those requirements that were identified in the charging scheme evaluation criteria (section 3.4.).

The chapter is divided into three key parts:

1. Trial set-up description (section 5.1.);
2. Simulation of the charging schemes (sections 5.2., 5.3., 5.4., 5.5. and 5.6.);
3. Evaluation and comparison of the charging schemes (section 5.7.).

Within part two (sections 5.2.4., 5.3.4., 5.4.1., 5.5.4. and 5.6.4.), a general remark is made to summarise the results of the experiments for each of the charging schemes.

5.1. TRIAL DESCRIPTION

Three traffic sources were used in the tests, representing the key service types that can cause the maximum difficulties in measurements and estimation due to the nature of their characteristics. These are as follows:

1. real-time VBR based source for multimedia services (ISABEL stream);
2. real-time VBR based source for video on demand services (MPEG stream);
3. real-time CBR based source.

UBR and ABR traffic is not supported by the existing schemes and the proposed scheme only requires cell count and minimum cell rate information to deduce charges for these traffic. Therefore simulations were not performed with regards to sources based on these transfer capabilities.

The ISABEL source is a "moving JPEG" video conference application that sends still JPEG pictures at a rate of 9 frames per second, each frame being 172 cells long.

Multiplexed with this is a voice-activated microphone transmitting 12 frames per second. The combined mean rate is 1.8Mbit/s.

The MPEG-I source is a trace (of up to 1 hour in length) from the Bond movie "Goldfinger" which has a mean rate of 0.6Mbit/s and frame length of 40ms, each frame consisting of a varying number of ATM cells.

The CBR source is a 2.048Mbit/s continuous traffic stream.

5.1.1. Input Source Characteristics

Figure 15 and Figure 16 illustrates the characteristic of the two sources (ISABEL and MPEG): the source rate distribution. The measurements for the graphs were obtained directly from the traffic sources before any buffering or multiplexing was carried out. The x-axis represents the bandwidth of the traffic source and the y-axis represents the probability of the source operating at the respective bandwidth. (The sum of the values at the y-axis equals to 1.)

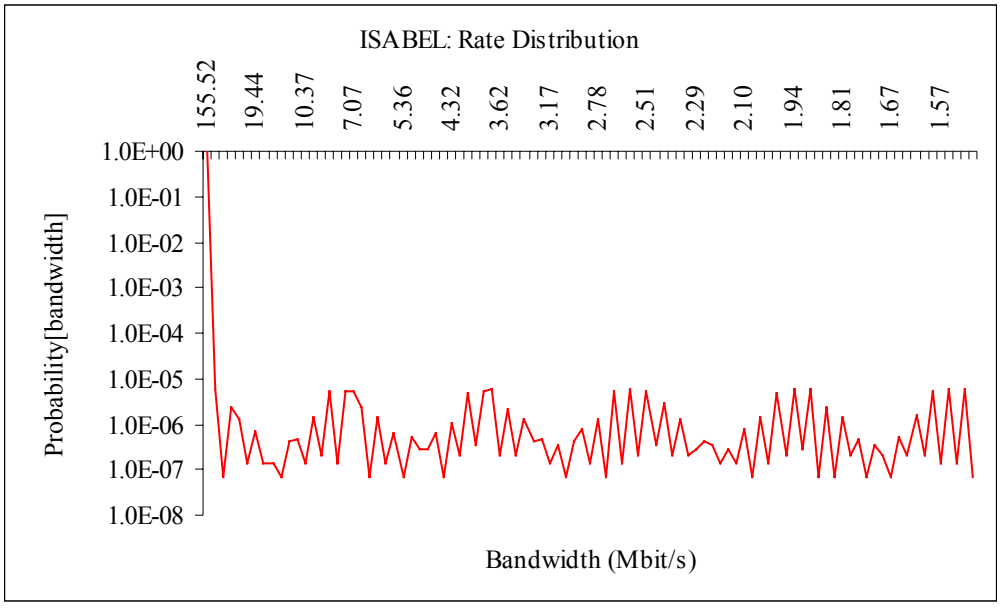

Figure 15: ISABEL source rate distribution.

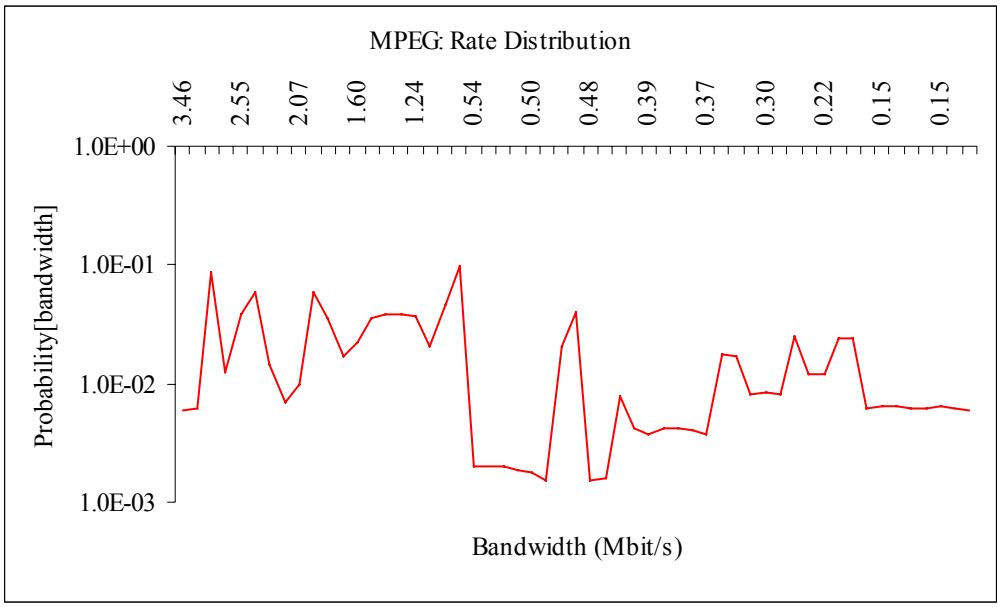

Figure 16: MPEG source rate distribution.

When these sources are multiplexed into a buffer of size 10 cells, service rate 155Mbit/s, and CLP requirement of 10^{-8}, the following results can be obtained (see section 4.5.5., Table 8 is for a service rate of 37Mbit/s and CLP of 10^{-4}):

- ISABEL source has
  1. mean bandwidth of 1.8Mbit/s
  2. peak bandwidth of 155Mbit/s
  3. effective bandwidth (calculated by simulations) of 2.8Mbit/s
- MPEG source has
  1. mean bandwidth of 0.6Mbit/s
  2. peak bandwidth of 3.5Mbit/s
  3. effective bandwidth (calculated by simulations) of 1.09Mbit/s
- CBR source has
  1. peak bandwidth of 2.048Mbit/s

5.1.2. Simulation Configuration

The switch configurations used in the simulations (for all the input traffic sources) are as shown in Figure 17 for the charging schemes proposed by Kelly, Lindberger and Botvich. For Griffiths' charging scheme the configuration is as shown in Figure 18, and Figure 19 depicts the configuration used for the QoS based charging scheme. In all cases, the buffers operate in FIFO mode with deterministic server.

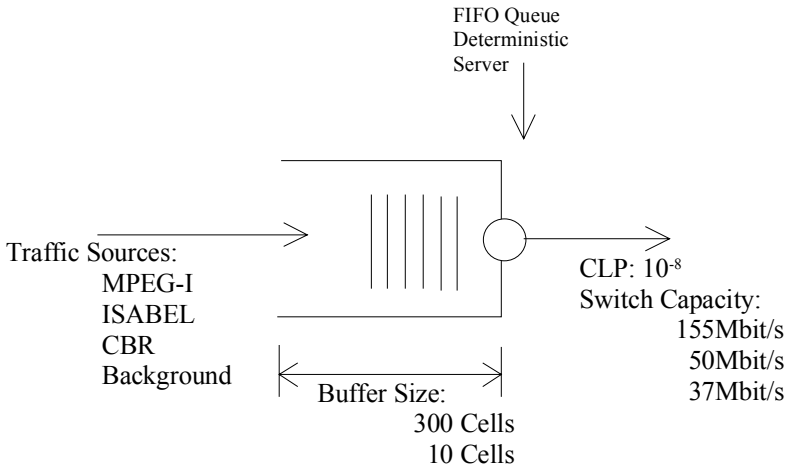

Figure 17: Trial configuration for the effective bandwidth based charging schemes.

With this configuration (Figure 17), two buffer sizes were used for experiments: 300 and 10 cells. The purpose of the second set of experiments was to compare the results from the simulations of the QoS based charging scheme, which used traffic shapers of ~290 cells and multiplexer of 10 cells (total of 300 cells). The buffers were also dimensioned to support a CLP of 10^{-8}.

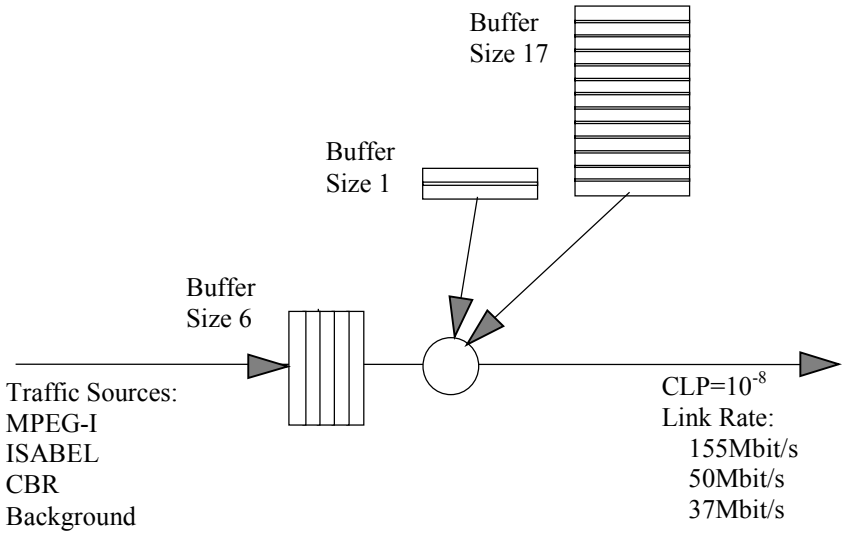

Figure 18: Trial configuration for the filter based charging scheme.

The filter used for the Griffiths scheme (Figure 18) was dimensioned according to the specification suggested by the author of the scheme: CLP of 10^{-8} and buffer sizes as shown in the diagram. This dimension applies to all the input traffic sources considered for the simulations reported here.

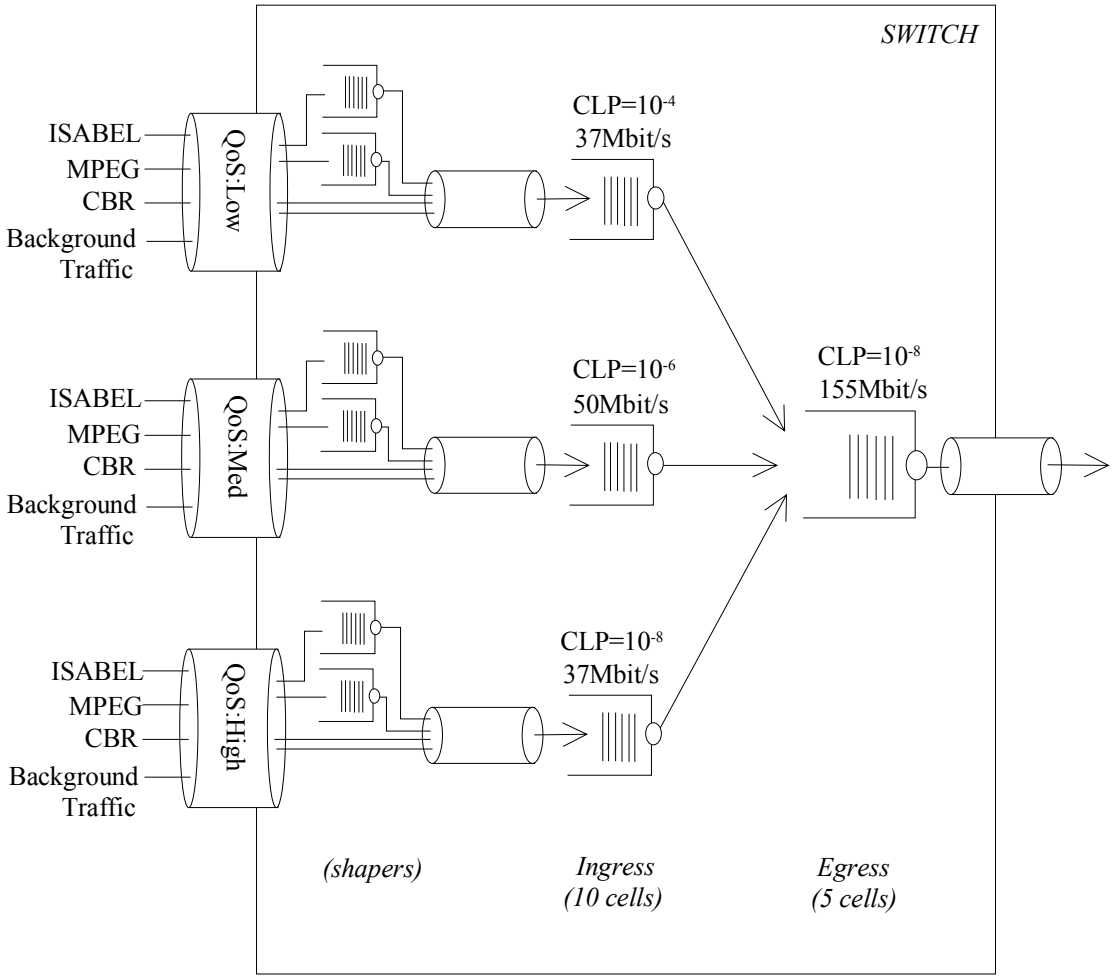

Figure 19: Trial configuration for the QoS based charging scheme.

For the QoS based charging scheme, the configuration (Figure 19) represents an access switch of capacity 155Mbit/s providing 3 QoS streams. Multiplexed into these streams were the input sources (ISABEL, MPEG and CBR) together with background traffic to load the switch to its maximum capacity. Background sources were also used in the simulations for the existing charging schemes. The streams were configured according to the theory proposed in chapter 4., giving the following set-up states:

1. Ingress buffers 10 cells, to support real-time VBR and CBR sources.
2. Egress buffer size of 5 cells, to support up to 5 streams of any QoS.
3. Shaper buffer sizes of 265 (ISABEL) and 236 (MPEG).
4. Bandwidth (switch capacity) distributed among the streams (according to assumed bandwidth distribution factors, $ingress\_bd_j$, of 25%, 33% and 25%) were 37Mbit/s, 50Mbit/s and 37Mbit/s. The remaining 36Mbit/s (from 155Mbit/s) of bandwidth was set aside for additional QoS streams.

The simulation tools used were:

1. YATS developed by the European ACTS project EXPERT.
2. Code written by the author for the effective bandwidth estimation for schemes by Kelly, Lindberger and Botvich.
3. Code written by the author of the fourth scheme (Griffiths) for the equivalent bandwidth calculations.

5.1.3. Trial Objectives

The qualitative evaluation of the existing charging schemes raised certain questions that must be answered by the simulation results; each scheme has its own set of issues that need clarification as described below.

Kelly's Scheme

Kelly's effective bandwidth algorithm requires that the source rate distribution be formulated online for charging purposes. For that the traffic streams are monitored (scanned for Sd seconds) every Si seconds (the scanning interval) to obtain a bandwidth, from which the source rate distribution curve is formulated. Figure 20 illustrates the process.

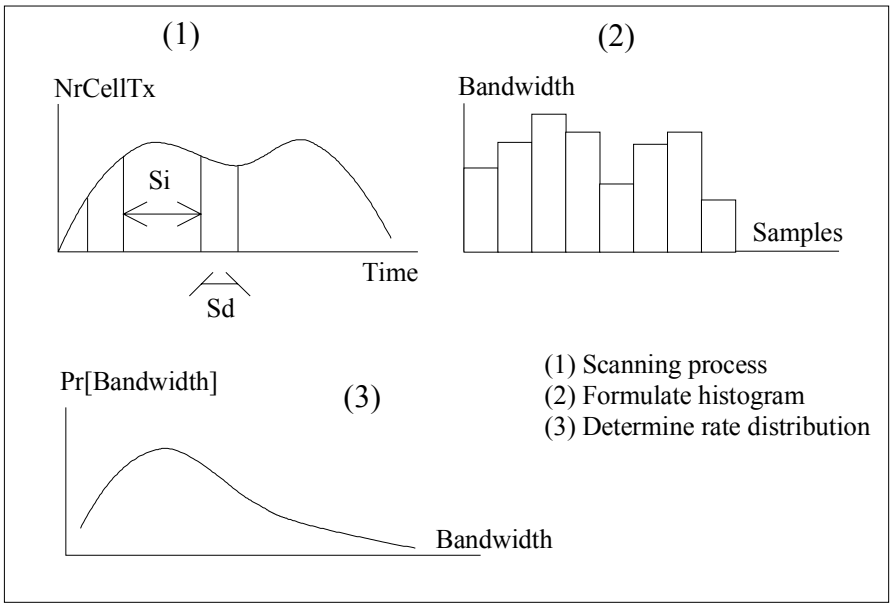

Figure 20: Scanning process for creation of source rate distribution.

Since the effective bandwidth that is evaluated directly affects charging, the following observations are required from the simulation results:

1. estimation of the effective bandwidth compared with the "real" value (see explanation below);
2. effect of altering the scanning parameters on the effective bandwidth estimation;
3. accuracy of the bandwidth estimation due to varying the duration of the connection;
4. the length of time it takes to evaluate the effective bandwidth;
5. effect on charges due to under or overestimation of the effective bandwidth.

The term "real" in bullet point (1) refers to the minimum bandwidth reservation judged (through simulations and tests with real network and live traffic) to be necessary in order to receive an acceptable level of pictures and sound quality from the sources under investigation. This was discussed in detail in section 4.5.5.

Lindberger's Scheme

Lindberger uses a similar approach to Kelly's (although the effective bandwidth algorithm is different) for calculation of charges. Therefore, the observations that are to be made for this scheme are identical to those of Kelly's except for the consequence of under or overestimation (there are no penalty mechanisms involved).

Botvich's Scheme

With Botvich, the focus remains on the calculation of the effective bandwidth using Kelly's algorithm since the k parameter used for charging is directly related to that. Hence the observations that are to be made for this scheme are identical to those of Kelly's except the consequence of under or over estimation (there are no penalty mechanisms involved).

Griffiths' Scheme

With Griffiths' scheme, it is necessary to obtain from the simulation results the answers to the questioning "how well does the filter perform and serve to provide accurate resource usage information for charging", and "how difficult is it to set-up"?

QoS Based Scheme

With this scheme, the points of interests that are to be observed from the simulation results are: what is:

1. the CTD introduced by the proposed switch configuration;
2. the CLP introduced by the proposed switch configuration;
3. resource requirement, β, evaluated for the purpose of charging.

5.1.4. Simulation Validation

The simulation tool, YATS, was developed by the University of Dresden, a partner in European ACTS project EXPERT (AC024). It has been validated by [BAU01]:

1. Checking the code at University of Dresden.
2. Checking the results of simulations against results from a test network using live traffic as part of the European project EXPERT.

The author also performed subjective tests using real network and live traffic to confirm that the conclusions reached from the experiments described in this chapter are valid. These tests were performed using MPEG-I traffic from the ORACLE Video Server for connections set-up over an APON access network in Fujitsu Telecommunications Europe Limited; see section 4.5.5.

In addition, the results obtained from the algorithms used to evaluate the effective bandwidth for schemes proposed by Kelly, Lindberger and Botvich were compared with results that were provided in [AAR01]. In these references the authors of the papers used different analytical and simulation tools to deduce the results.

Algorithms used to compute the equivalent bandwidth for the Griffiths' scheme were also validated using the EXPERT testbed, by the author of the scheme [GRIF03].

Furthermore, the code for the bandwidth estimation algorithms was checked line by line (by the author).

Finally, for every experiment, at least 10 simulations were performed to check the consistency of the results. Each graph presented in this chapter is outcome of a single simulation that represents the typical result for that experiment. Note also that for the

purpose of presentation and discussion the curves in all the graphs presented have been smoothed; Figure 21 shows one example of the differences between the actual results and the curve drawn from it.

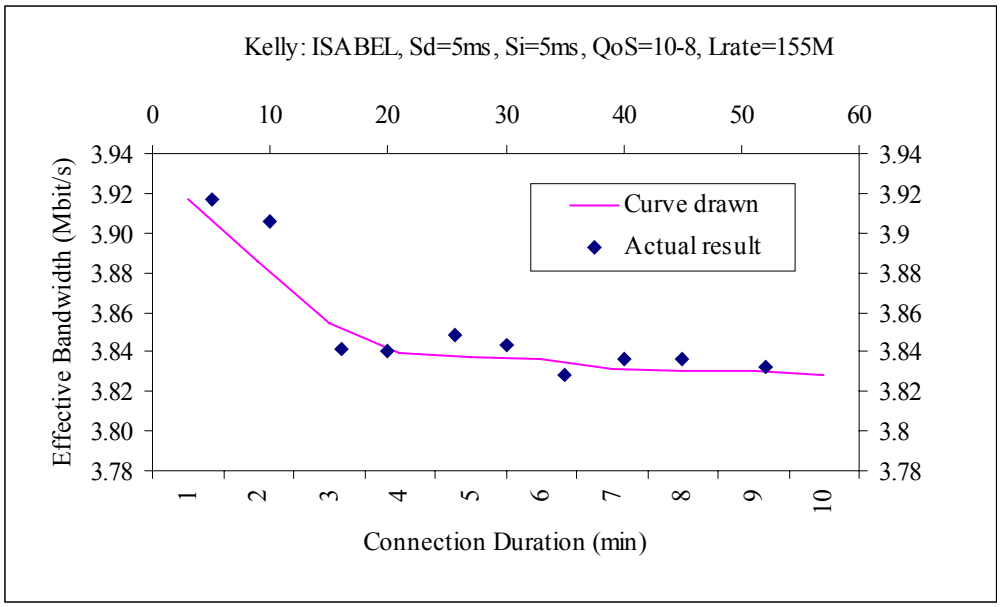

Figure 21: Error margin in the simulation results due to curve fitting.

5.2. Results of Simulating Kelly's Charging Scheme

The results from the simulations of each of the charging schemes are presented in nine separate graphs: four for the ISABEL experiments; four for the MPEG; and one to show the effect of over or under estimation of the declared bandwidth.

For the ISABEL and MPEG sources, each graph represents two experiments; one with a buffer size of 300 cells and the other with a buffer size of 10 cells.

5.2.1. Results from Simulations using the ISABEL Source

Figure 22 shows the effect of varying the scanning duration; the rest of the parameters were fixed at the following reference points:

- scanning interval = 0 (continuous scanning);
- link rate (switch capacity) = 155Mbit/s;
- cell loss probability = 10^{-8};
- duration of the connection = 60 minutes.

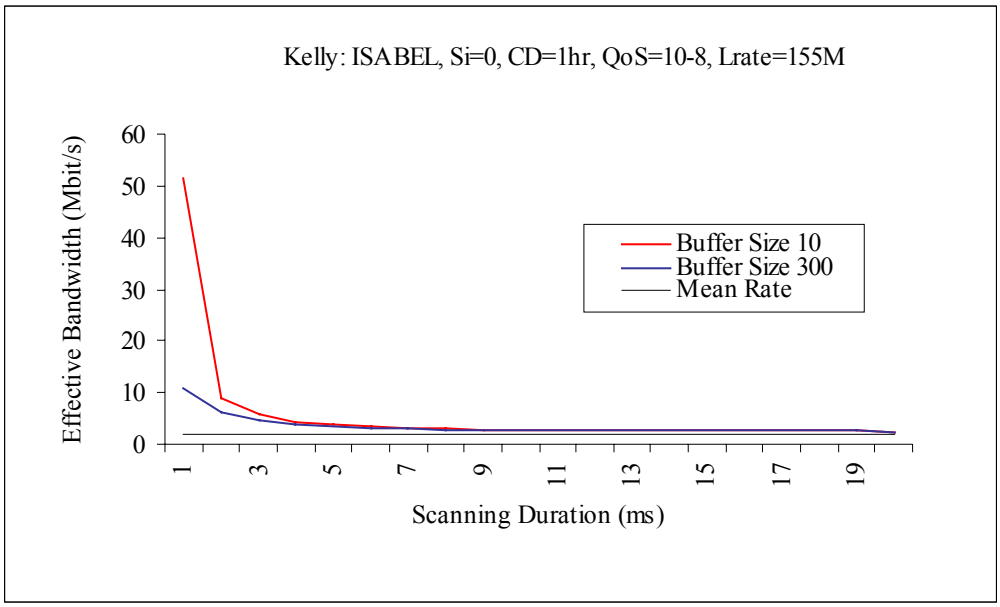

Figure 22: Kelly - effect of varying scanning duration for ISABEL source.

It can be seen that the effective bandwidth approximates to slightly above the mean rate as the scanning duration is increased. On the other hand, reducing the scanning duration down to a single slot produces an effective bandwidth close to the peak rate. The network operator must decide on a suitable value for the scanning duration, which reflects accurately the actual bandwidth requirement that can be supported by the network and which can be used as a fair basis for charging.

Selecting a scanning duration of 5ms as the reference point for the rest of the experiments, the scanning interval was varied to observe its effect on the evaluation of the effective bandwidth; Figure 23 depicts the result. Note that, at this point (and for the purpose of the next experiment) the choice of the scanning duration is arbitrary since the objective is only to observe what effect the variation of the scanning interval has on the effective bandwidth estimation.

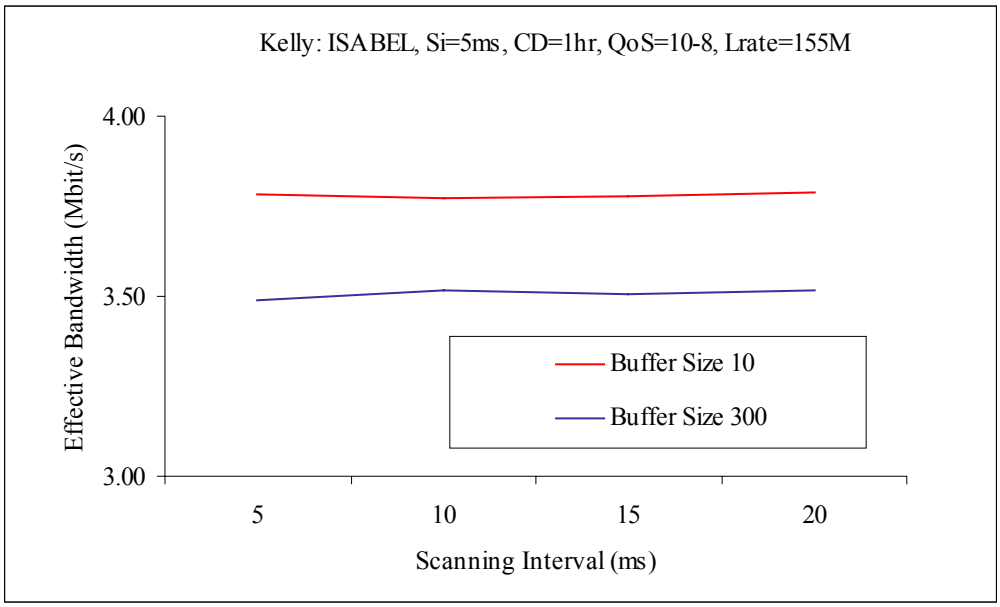

Figure 23: Kelly - effect of varying scanning interval for ISABEL source.

Little or no effect could be observed on the effective bandwidth calculation due to variation of the scanning interval (Figure 23). This is because, provided that the connection duration is long enough, a sufficient number of samples are collected to evaluate the effective bandwidth. For short connection duration, the scanning interval may have to be reduced to (almost) a single time slot to ensure that the sample number is large enough, or else the bandwidth estimated will not be reliable. Note however that the operators will not know how long a connection will last at the time of the connection set-up.

The effect of the connection duration is well demonstrated in the next experiment. Setting the scanning interval to 5ms, the duration of the connection was varied; one result of this is shown in Figure 24. Note that, at this point (and for the purpose of the next experiment) the choice of the scanning interval is arbitrary since the objective is only to observe what effect the variation of the connection duration has on the effective bandwidth estimation.

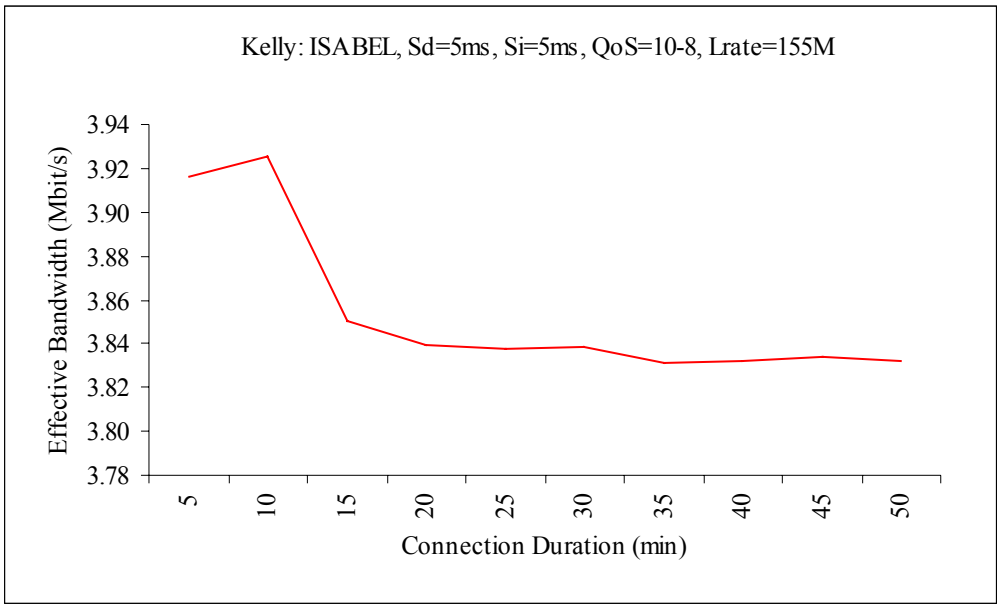

Figure 24: Kelly - effect of varying connection duration for ISABEL source.

It can be observed that the effective bandwidth may vary unpredictably if the duration of the connection is too short. In the case of ISABEL, with scanning interval and scanning duration set to 5ms, the connection duration must be at least 15 minutes in order to achieve a reasonable estimation of the effective bandwidth.

Kelly's effective bandwidth algorithm also takes into account the switch capacity (link rate). Figure 25 shows result of the variation in the link rate.

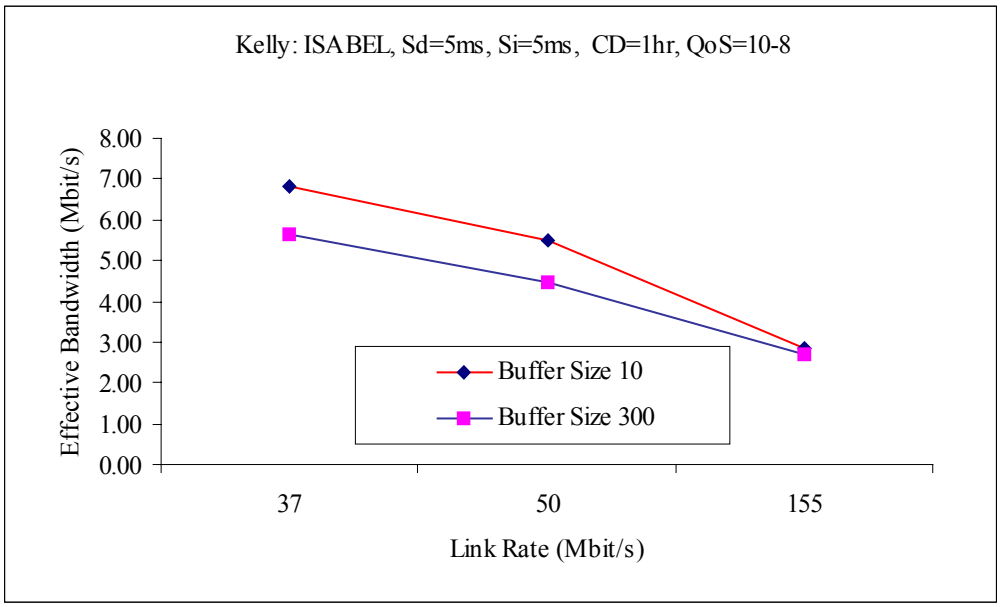

Figure 25: Kelly - effect of varying link rate for ISABEL source.

Notice that the lower the link rate for the connection, the higher is the effective bandwidth estimated.

5.2.2. Results from Simulations using the MPEG Source

This source has the property that the information is sent in a frame consisting of a variable number of cells every 40ms. As before, the first experiment was for the observation of the effect of varying the scanning duration given the following set-up:

- scanning interval = 0 (continuous scanning);
- link rate = 155Mbit/s;
- cell loss probability = 10^{-8};
- connection duration = 60 minutes.

The result is shown in Figure 26.

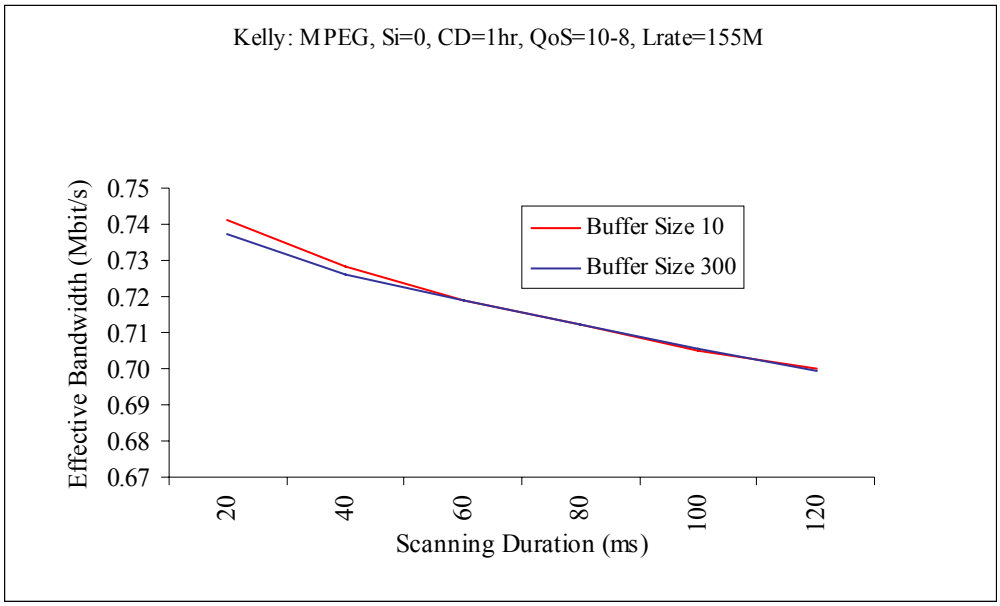

Figure 26: Kelly - effect of varying scanning duration for MPEG source.

As with the previous results, increasing the scanning duration has the effect of moving the bandwidth closer to the mean bit rate (0.6Mbit/s) of the source and decreasing it move it closer to the peak rate. Again, the network operator must decide on what is the most appropriate setting for this service.

Note that, at this point (and for the purpose of the next experiment) the choice of the scanning duration is arbitrary since the objective is only to observe what effect the variation of the scanning interval has on the effective bandwidth estimation.

Setting the scanning duration to 40ms, Figure 27 shows the effect of varying the scanning interval. As for ISABEL, this has no effect on the result, provided that the connection duration is long enough to collect a sufficient number of samples. If not, then the bandwidth estimation will not be reliable.

Note that, at this point (and for the purpose of the next experiment) the choice of the scanning interval is arbitrary since the objective is only to observe what effect the variation of the connection duration has on the effective bandwidth estimation.

Hence the scanning interval of a single slot (si=0) was set for the remainder of the experiments.

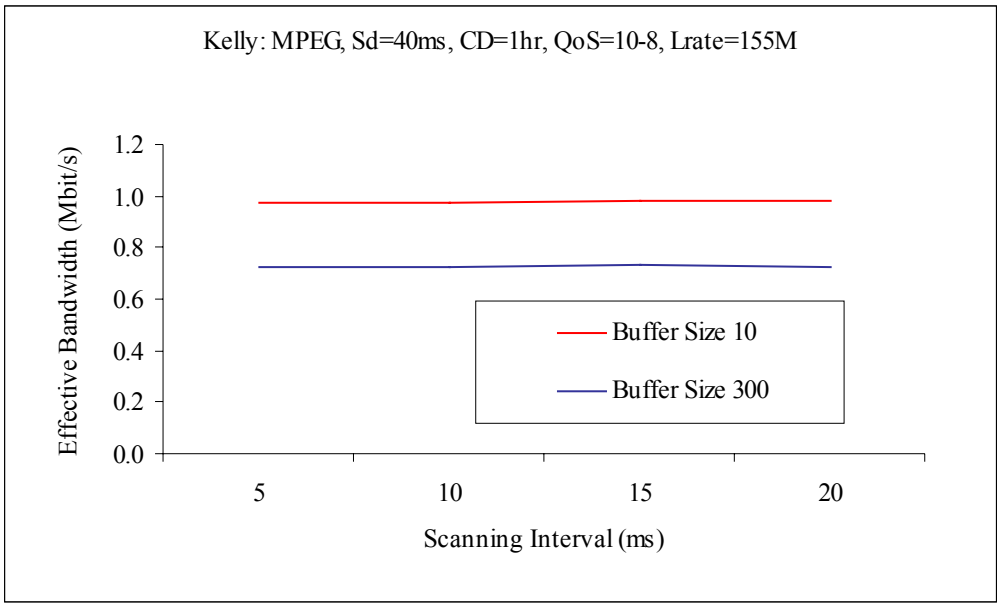

Figure 27: Kelly - effect of varying scanning interval for MPEG source.

Having set the scanning duration and scanning interval, the effect of varying the connection duration was investigated; Figure 28 shows the outcome. It can be seen that a connection duration of 25 minutes or above at this rate of scanning is sufficient for evaluating a consistent set of effective bandwidth values.

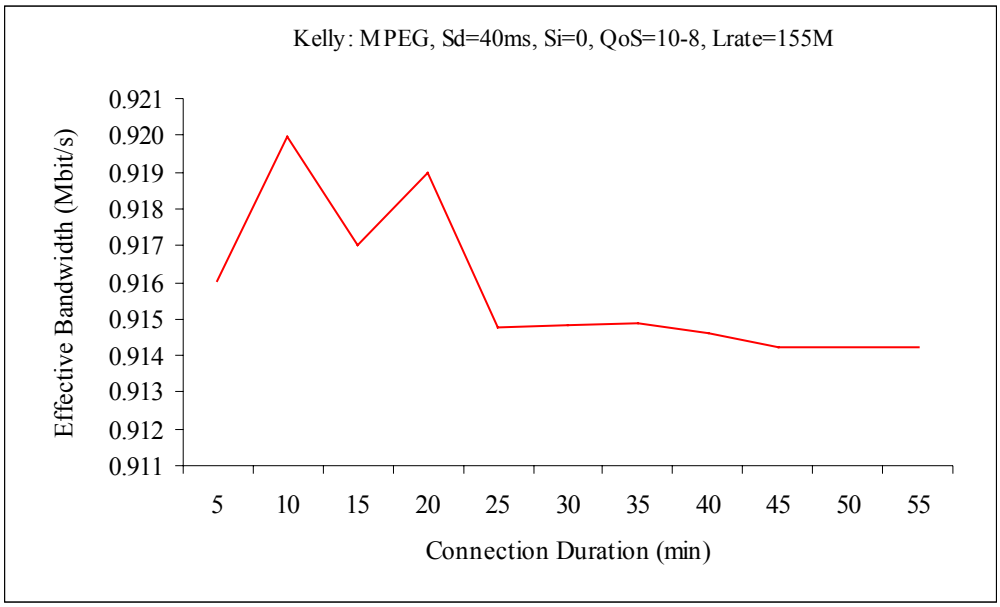

Figure 28: Kelly - effect of varying connection duration for MPEG source.

Looking at the effect of the link rate on the effective bandwidth estimation, the same observations as for the ISABEL experiments are apparent. This is depicted in Figure 29.

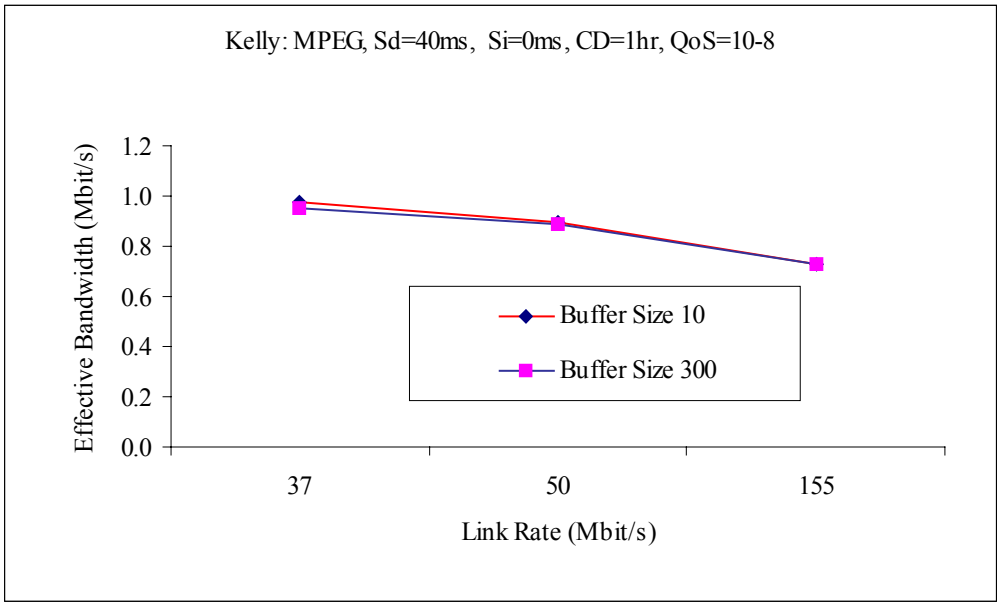

Figure 29: Kelly - effect of varying link rate for MPEG source.

5.2.3. Result of the CBR source

There are no effective bandwidth calculations for the CBR source; a peak rate of 2.048Mbit/s is taken as the resource allocation and charging parameter.

5.2.4. General Remarks on the Simulation Results of the Kelly's Scheme

The difference between the results of the simulations with multiplexer buffer size of 10 cells and multiplexer buffer size of 300 is not very significant. As far as the algorithm is concerned any buffer size lower than 500 cells is considered to be in the small-buffer category.

The most significant component in the scanning process is the identification of the scanning duration. The effective bandwidth will vary between the peak rate and almost the mean rate of the traffic source depending on the scanning duration selected. Furthermore, the correct scanning duration for the traffic sources will differ between different source type. What is the correct value for any source is debatable, particularly when charges are involved.

The duration of the connection is also very important. If a customer does not stay on-line for a sufficiently long period of time (15 minutes in some cases) the effective bandwidth estimated might not be accurate; the estimation and therefore charges will be either too high or too low.

Customers must be aware that the effective bandwidth estimated by the operator may not always be accurate or closely match the declared value; the estimation is highly dependent on the scanning parameters set and the duration of the connection. Figure 30 shows the effect on the charge due to customers failing to declare the effective bandwidth equal to that which the network operator will estimate when using the Kelly's charging scheme.

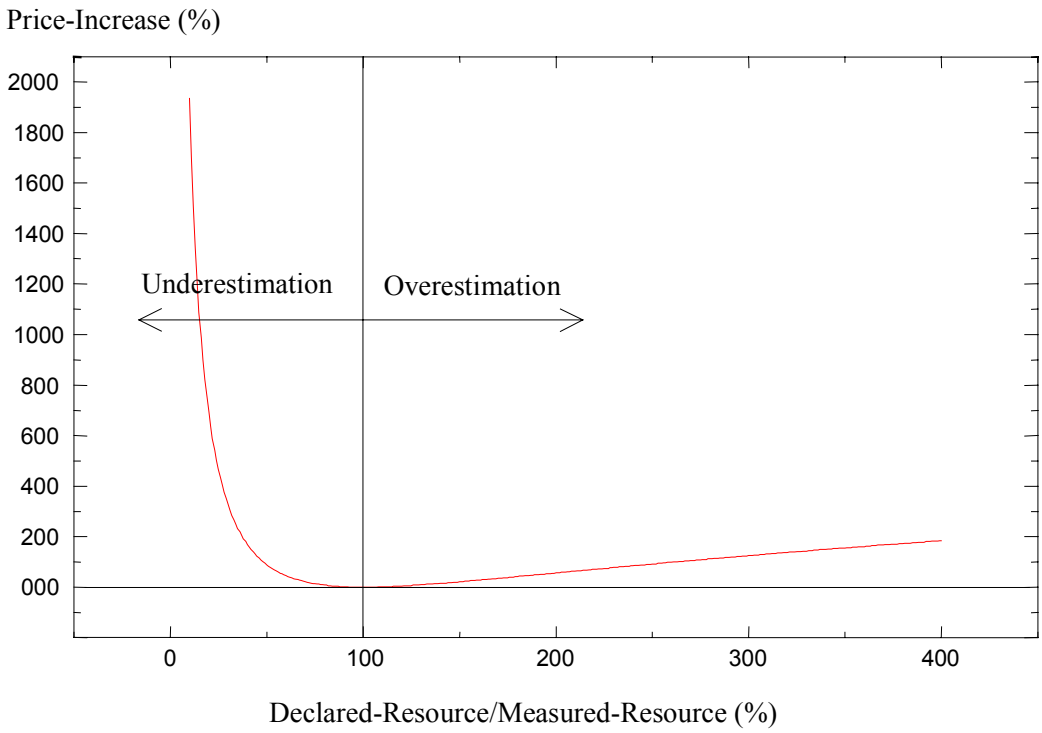

Figure 30: Effect of effective bandwidth under estimation.

Furthermore, the link rate over which a connection is made can also effect the bandwidth estimation (Figure 25 and Figure 29). For example, two customers running the same service but using ATM-25 and STM-1 interfaces will be charged differently.

There were a number of other difficulties that were encountered during the experiments. For example, it was found that for a large number of samples, the algorithm took a very long time to compute; in the order of 30 minutes for the ISABEL source. At best the speed was calculated to be 3 minutes on a SUN SPARC ULTRA 20 workstation. In certain cases, such as the link rate of 37Mbit/s and cell loss probability of 10^{-8} the algorithm could never be completed; the bandwidth estimated is too close to the actual link rate and the algorithm reaches a permanent loop.

One further issue should be noted at this point. The effective bandwidth estimated by the algorithm cannot be used to allocate resources without use of traffic shapers, as was described in section 4.5.5. This is true even if the bandwidth estimation closely matches that which is determined through experimentation.

5.3. RESULTS OF SIMULATING LINDBERGER'S CHARGING SCHEME

With Lindberger's method, the charge is also dependent on the effective bandwidth and therefore the simulations were similar to those that were carried out for Kelly's. The technique used in the bandwidth estimation is the same as that adopted by Kelly in that the rate distribution for the traffic streams are re-created by scanning. However, with this approach, buffer size is not included in the calculation. Instead, the standard deviation (σ^2) is evaluated from the source rate distribution using the assumptions that the source is either ON/OFF or a statistically distributed type. Hence, with this experiment each graph contains the results of two different simulations; one assumes that the source is an ON/OFF type and the other assumes it to be statistically distributed.

5.3.1. Results from Simulations using ISABEL Source

In order to observe the effect of varying the scanning duration the following parameters were set as a reference point:

- cell loss probability = 10^{-8};
- link rate = 155Mbit/s;
- duration of the connection = 60 minutes;
- scanning interval = 2.5ms.

Figure 31 shows the result of the simulations. As with Kelly, the scanning duration does have a significant impact on the size of the effective bandwidth estimated, tending to the mean bit rate (1.8Mbit/s) as the duration increases and to the instantaneous bandwidth (which can be anything from zero to the peak) as the duration decreases. A suitable value has to be selected by the network operator to represent the actual bandwidth requirement.

Note that, at this point (and for the purpose of the next experiment) the choice of the scanning duration is arbitrary since the objective is only to observe what effect the variation of the scanning interval has on the effective bandwidth estimation.

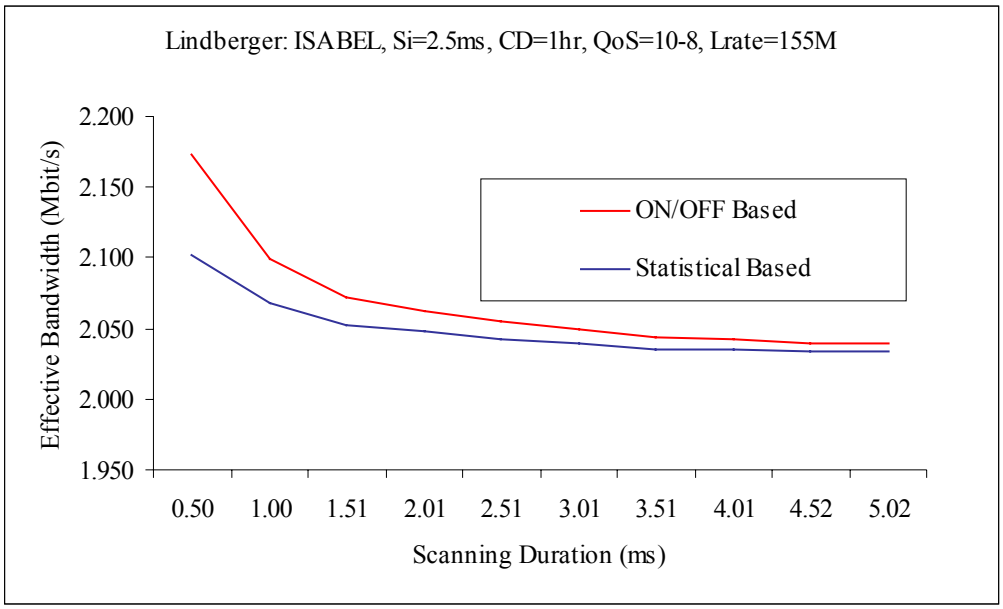

Figure 31: Lindberger - effect of varying scanning duration for ISABEL source.

Setting the scanning duration to 2.5ms for the remainder of the simulations, Figure 32 depicts the result of varying the scanning interval. Variation of the scanning interval has little effect on the outcome due to the fact that provided the connection duration is long enough for collecting a sufficient number of samples the effective bandwidth calculation is predictable. For a very short connection duration, the scanning interval will have to be reduced to (almost) single time slots. Note, however, that the operator will not know the duration of the connection at set-up.

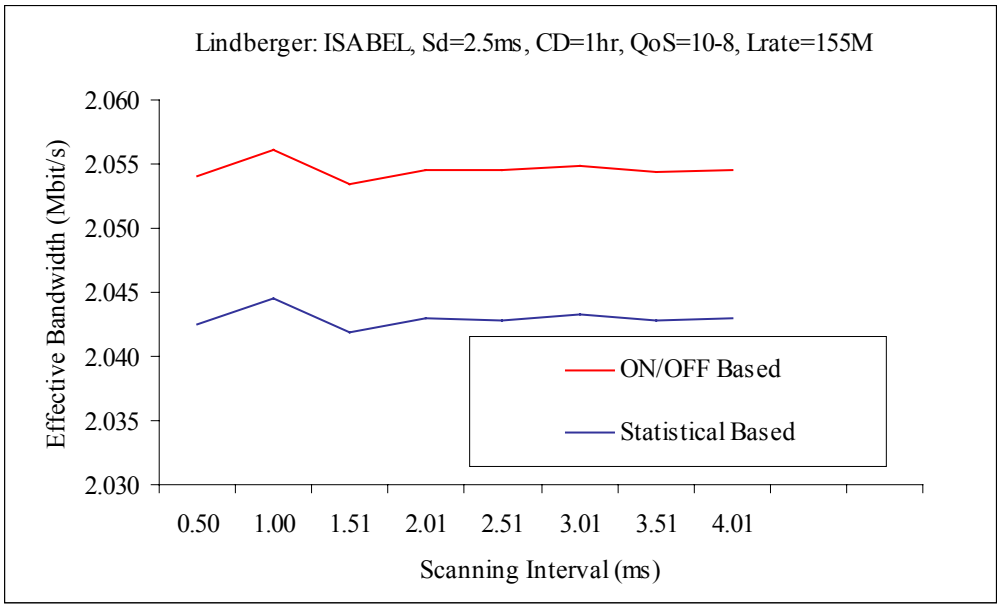

Figure 32: Lindberger - effect of varying scanning Interval for ISABEL source.

Note that, at this point (and for the purpose of the next experiment) the choice of the scanning interval is arbitrary since the objective is only to observe what effect the variation of the connection duration has on the effective bandwidth estimation. Let the scanning interval be set to 2.5ms for the remainder of the simulations.

Moving onto the next parameter, the effect of varying the duration of the connection is shown in Figure 33. It can be observed that the effective bandwidth varies unpredictably if the duration is too short. In the case of ISABEL, with a scanning interval of 2.5ms, this has to be at least 30 minutes.

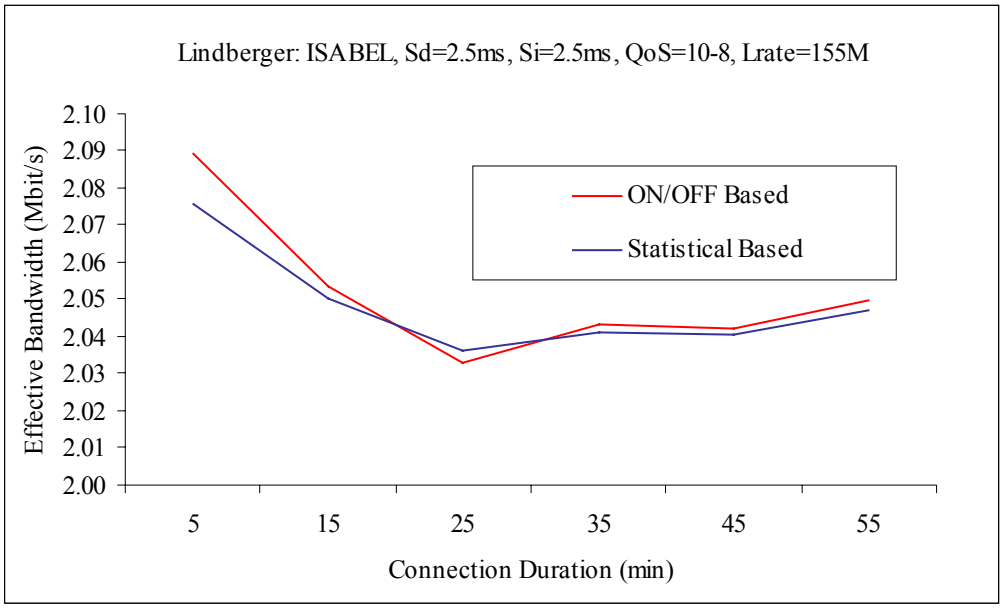

Figure 33: Lindberger - effect of varying connection duration for ISABEL source.

Lindberger's scheme also takes into account the link rate. Figure 34 shows the effective bandwidth calculated given this variation.

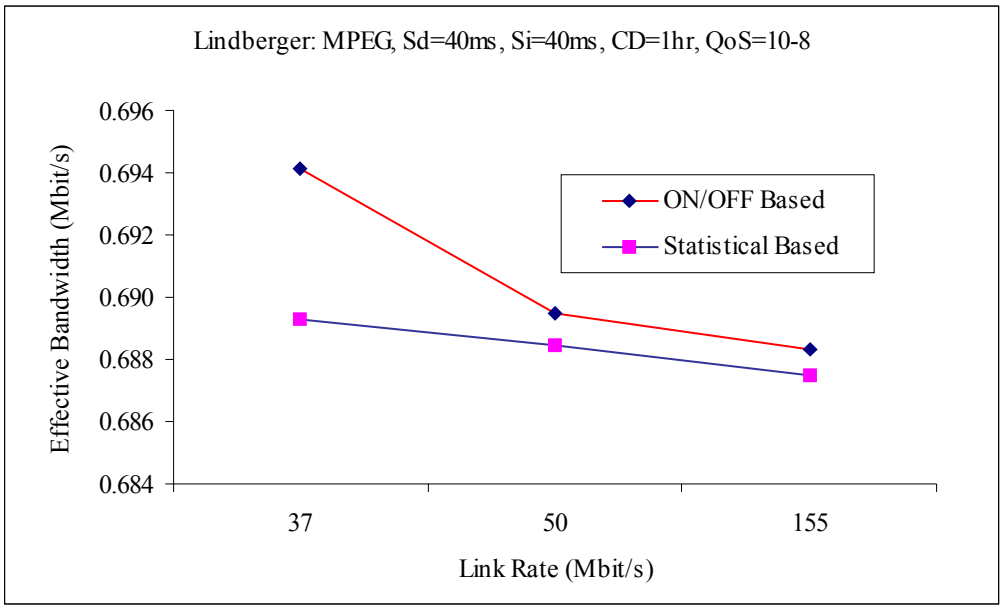

Figure 34: Lindberger - effect of varying link rate for ISABEL source.

Notice that the lower the link rate the higher is the effective bandwidth calculated.

5.3.2. Results from Simulations using the MPEG Source

This source has the property that all the information is sent in a frame of a variable number of cells every 40ms. The first set of simulations were carried out to determine the effect of varying the scanning duration with the rest of the parameters set at:

- scanning interval = 0;
- link rate = 155Mbit/s;
- cell loss probability = 10^{-8};
- duration of connection = 60 minutes;

The result of the simulations is shown in Figure 35.

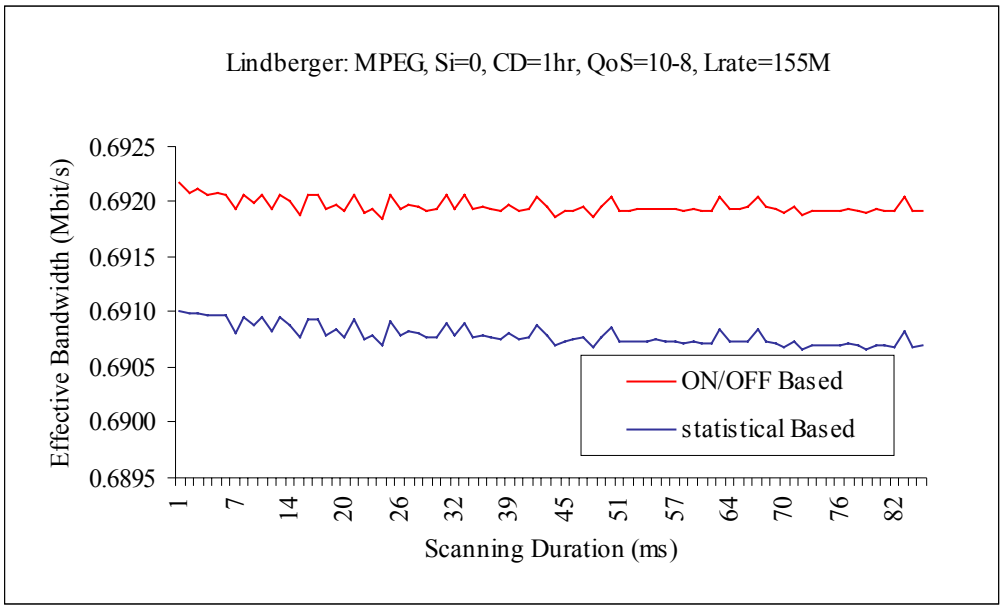

Figure 35: Lindberger - effect of varying scanning duration for MPEG source.

Notice that the scanning duration has little effect on the effective bandwidth calculated. The explanation is that the connection duration is sufficiently long (1 hour) to sample the entire range of source emission rate and thereby reproduce the rate distribution accurately. The scanning duration is set to 40ms for the remainder of the experiments with the MPEG source.

Note that, at this point (and for the purpose of the next experiment) the choice of the scanning duration is arbitrary since the objective is only to observe what effect the variation of the scanning interval has on the effective bandwidth estimation.

Figure 36 shows the effect of varying the scanning interval: again this has little effect on the outcome of the simulations.

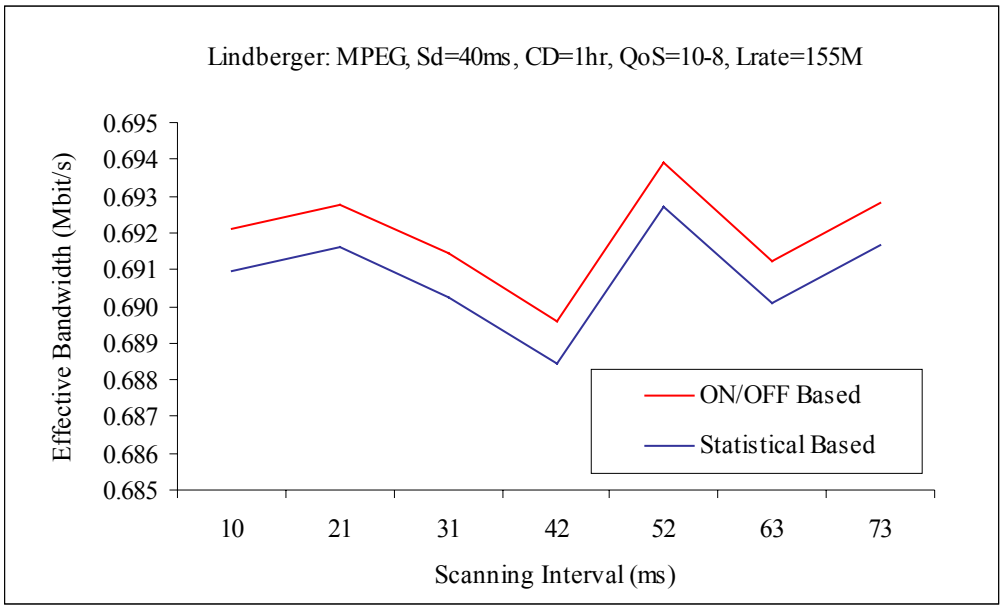

Figure 36: Lindberger - effect of varying scanning interval for MPEG source.

Setting the scanning interval to a single frame (40ms) to speed up the simulations (a frame is scanned every other frame) the result of varying the connection duration is shown in Figure 37.

Note that, at this point (and for the purpose of the next experiment) the choice of the scanning interval is arbitrary since the objective is only to observe what effect the variation of the connection duration has on the effective bandwidth estimation.

It can be seen that a connection duration of 40 minutes or more at the set rate of scanning duration and scanning interval is necessary for evaluating a consistent set of effective bandwidth values.

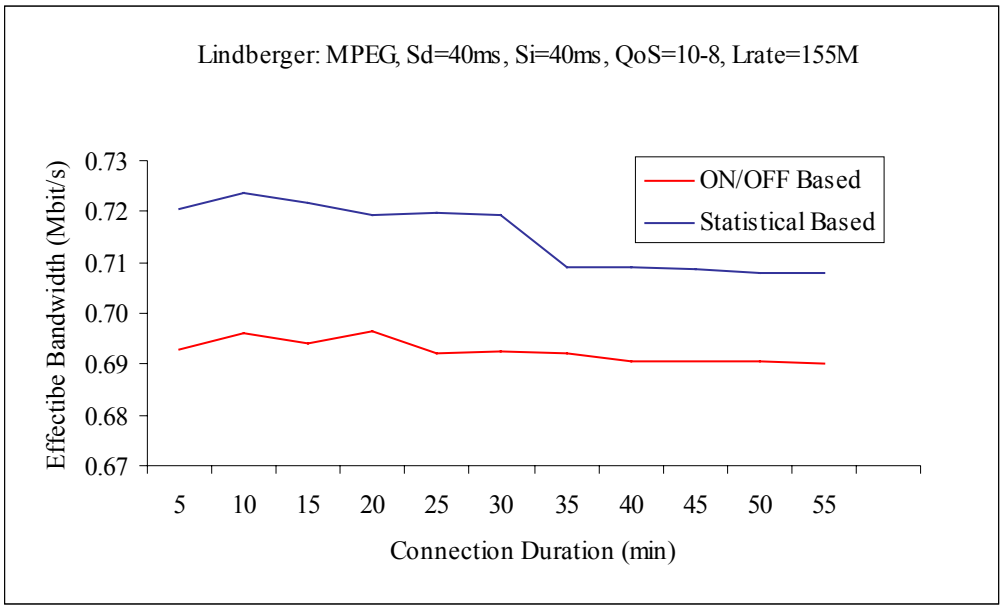

Figure 37: Lindberger - effect of varying connection duration for MPEG source.

Looking at the effect of the link rate, the same observation as for the ISABEL experiments can be observed; see Figure 38.

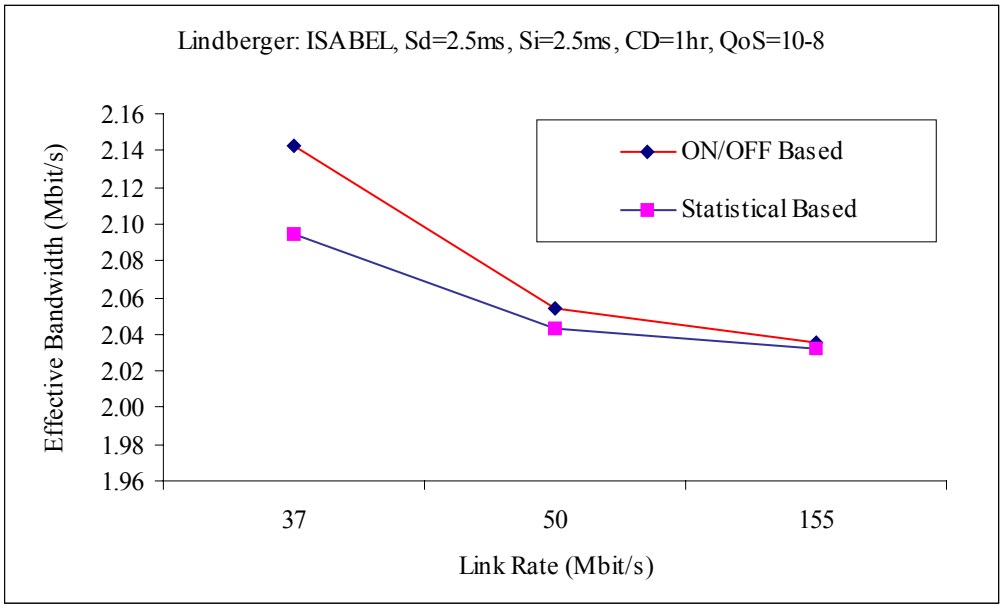

Figure 38: Lindberger - effect of varying link rate for MPEG source.

5.3.3. Result of the CBR Source

There is no need to calculate an effective bandwidth for the CBR source as its requirement is for a peak bit rate of 2.048Mbit/s and this is taken as the resource allocation and charging parameter.

5.3.4. General Remarks on Simulation Results of the Lindberger's Scheme

The difference between the effective bandwidth values estimated by assuming that the source is of type ON/OFF or statistical can be significant and the network operators must ensure that their assumption is accurate. Given the traffic characteristics it is safe to assume that the ISABEL is an ON/OFF type traffic while the MPEG could be viewed both ways.

The most significant component in the scanning process is the identification of the scanning duration. The effective bandwidth will vary between the peak rate and the mean rate of the traffic source depending on the scanning duration selected. Furthermore, the correct scanning duration for the traffic sources will differ between different source types. What is the correct value for any source is debatable, particularly when charges are involved.

The duration of the connection is also very important. If a customer does not stay on-line for a sufficiently long period of time (40 minutes in some cases) the effective bandwidth estimated might not be accurate; the estimation and therefore charges will be either too high or too low.

As with Lindberger's scheme, customer must be aware that the effective bandwidth estimated by the operator might not be what they expected; the estimation is highly dependent on the scanning parameters (used to monitor the traffic) and the duration of the connection. The bandwidth calculated is again dependent on the link rate over which a connection is run (Figure 34 and Figure 38). So for example, two customers running the same service but using STM-1 and ATM-25 interfaces will be charged differently.

One further issue should be noted at this point. In some cases, the effective bandwidth estimated by the algorithm cannot be used to allocate resources without use of traffic shapers, as was described in section 4.5.5. This is true even if the bandwidth estimation closely matches that which is determined through experimentation.

5.4. Results of Simulating Botvich's Charging Scheme

With this scheme, Kelly's effective bandwidth algorithm is used together with the mean rates of the connections. The mean rates for the two sources, ISABEL and MPEG, were 1.8Mbit/s and 0.6Mbit/s respectively. These values do not change and the parameter, k, is

$$k = \frac{\beta(z)}{m}$$

which makes k equal to $1.67*\beta(z)$ and $0.56*\beta(z)$ for the MPEG and the ISABEL sources used in the simulations. $\beta(z)$ is Kelly's effective bandwidth (see section 3.5.1.). All the observations on the simulations for Kelly also apply to this scheme. Hence, the graphs depicted in Figure 22 to Figure 29 are identical to those achieved for this scheme except that the y-axis now represents the k parameter instead of the $\beta(z)$ value.

5.4.1. General Remarks on Simulation Results of the Botvich Scheme

The difference between this scheme and Kelly's is that the k parameter is only evaluated once every x days/hours/minutes, off-line. From then on, the mean rate declared by the customer is used via $k*mean$ to allocate resource and make charges. Inaccurate declaration of resource requirement (incorrect mean rate declared by the customers or incorrect estimation of k) will result in loss of cells by the UPC or reserving more bandwidth than necessary and thereby overpaying.

Customers will have to be informed of the new k value that is updated periodically since this will affect the amount of resource they ask for in the future use of the service, and therefore affect the charge.

This scheme inherits all the problems associated with Kelly's effective bandwidth algorithm (k is evaluated using Kelly's approach): problems with scanning duration, connection duration and the link rate. To demonstrate this, the k parameter evaluated for the ISABEL source is shown in the graphs below.

Figure 39 shows the effect of varying the scanning duration. Notice that the value of k decreases to a minimal value as scanning duration is increased. On the other hand, reducing the scanning duration down to a single slot increases the k parameter rapidly. The network operator must decide on a suitable value for the scanning duration, which reflects accurately on the actual requirement, that can be supported by the network and which can be used to base the charges fairly.

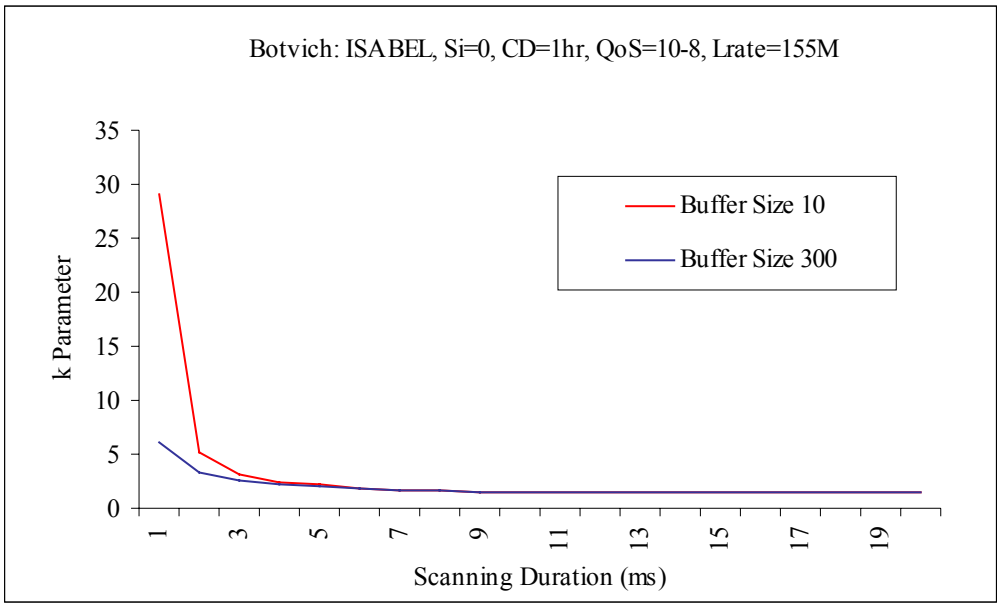

Figure 39: Botvich - effect of varying scanning duration for ISABEL source.

Figure 40 depicts the result of varying the connection duration, which must be at least 40 minutes to achieve a reliable estimation of the *k* parameter.

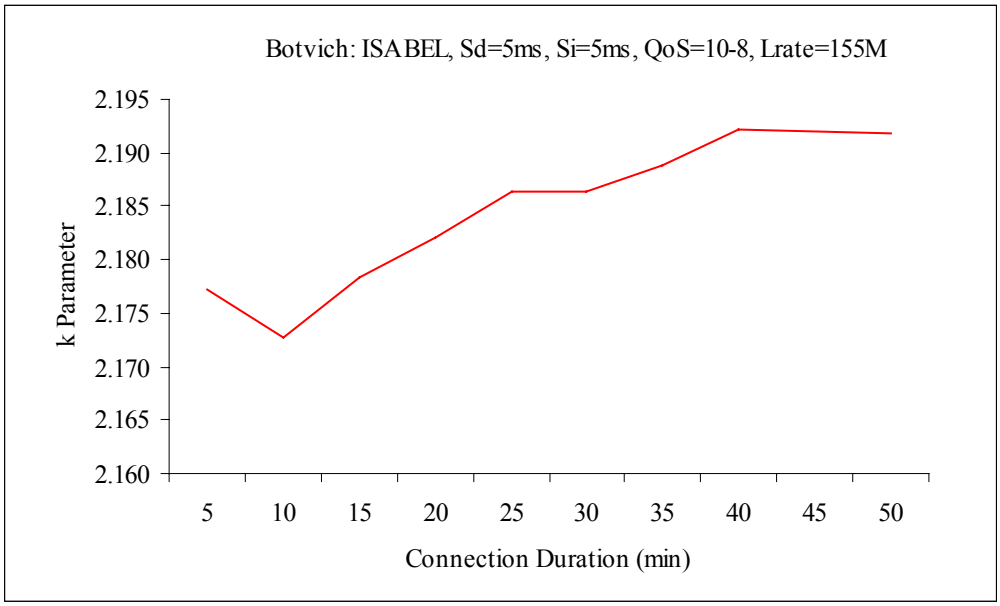

Figure 40: Botvich - effect of varying connection duration for ISABEL source.

Figure 41 shows the *k* parameter calculated given the variation in the link rate. Notice that the lower the link rate, over which the connection is supported, the higher is the *k* parameter estimated.

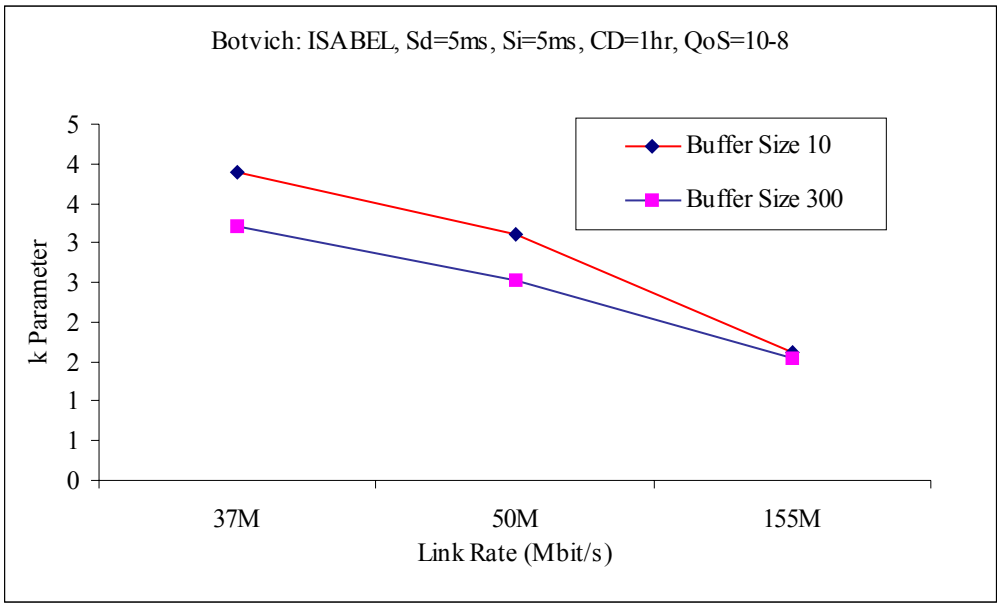

Figure 41: Botvich - effect of varying link rate for ISABEL source.

5.5. Results of Simulating Griffiths' Charging Scheme

With this scheme, customers request a particular bandwidth, called the design rate (*dr*), and the network implements a filter that ensures that the output of the filter is an equivalent bandwidth (*eb*) which is, at worst, not more than 1.5 times the design rate. If the customer violates the contract and increase the *dr* during the connection without re-negotiation with the network, the filter will discard cells. Customers pay for the *eb*.

Simulations were carried out to determine the *eb* for varying *dr* for the three sources, ISABEL, MPEG and CBR.

5.5.1. Result for the ISABEL Source

Figure 42 shows the result of the simulations performed with the ISABEL source. The x-axis shows the design rate, *dr* (input of the filter). The y-axis shows the equivalent bandwidth, *eb* (output of the filter). It can be seen that the output of the filter follows the input closely, but as the input rate increases, the curve becomes non-linear and the equivalent bandwidth (for which charges are made) becomes greater than the design rate. However, the increase remains below the 1.5 times design rate, as stated by the author of the scheme. Also, there is some variation in the outcome depending on the link rate (switch capacity).

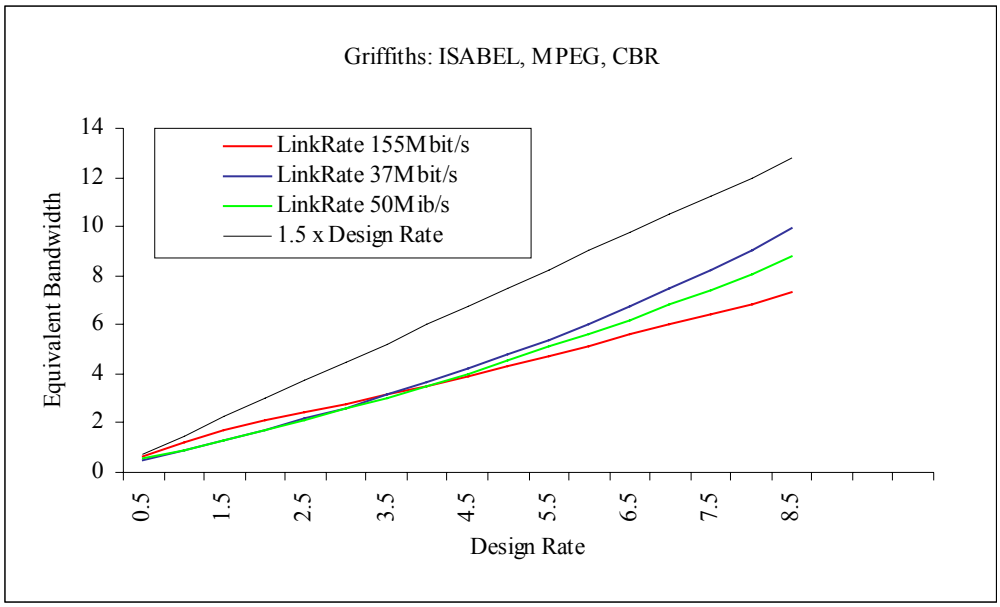

Figure 42: Griffiths - effect of varying design rate and link rate.

5.5.2. Result of the MPEG Source

The result of the MPEG source was identical to that achieved for the ISABEL source, Figure 42.

5.5.3. Result of the CBR Source

For the CBR source the peak rate of 2.048Mbit/s is taken as the resource allocation and charging parameter.

5.5.4. General Remarks on Simulation Results of the Griffiths' Schemes

When the filter rate (equivalent bandwidth) goes above 1.5 times the design rate (input rate) the filter is re-dimensioned and the buffers are resized. The recalculation involves a trial and error approach where the buffer sizes are "guessed" and the filter rate is checked. The whole process is re-visited if the outcome is not the desired value (i.e., until the equivalent bandwidth is less than or equals to 1.5 time the design rate).

Note that Figure 42 does not make clear what design rate the customer should choose.

5.6. RESULTS OF SIMULATING QOS BASED CHARGING SCHEME

For the QoS based charging scheme prices are based on the resource requirement β through the expressions in (Equ 30) and (Equ 31). The parameter β is equated to the peak cell rate for the CBR based services, volume of cells for the UBR based services and both the minimum cell rate and volume of cells transmitted for the ABR based services. For the VBR based services, β represents the shaper leak rate where the shaper is dimensioned using the equations (Equ 57) and (Equ 58). The simulations for QoS based charging scheme were performed for two objectives:

1. To dimension the shapers and determine β for the sources ISABEL and MPEG.
2. To confirm that the configuration shown in Figure 19 is such that:
    - the bandwidth distribution (155Mbit/s switch capacity distributed among the ingress buffers as shown in Figure 19) is feasible;
    - the overall QoS requested by the individual connections are conformed to;
    - the β determined is "correct" and "practical" for all the sources.

The term "correct" and "practical" refers to the minimum bandwidth reservation judged (through simulations and tests with real network and live traffic) to be necessary in order to receive an acceptable level of pictures and sound quality from the sources under investigation. This was discussed in detail in section 4.5.5.

The second objective was proven through a multitude of prolonged (10^{21} time slots) simulations. The result of one such simulation is shown in the database presented in Appendix A, which shows the exact status of each of the components existing for each of the QoS stream; i.e., β, CLP and CTD for each of the three streams and the connections that are supported by them.

In the following sections, results of the simulations carried out to determine β are shown. However, before moving onto this, a number of calculations are necessary, given the configuration and the scenario that is simulated, in order to prepare for the dimensioning work of the shapers and determination of the β parameter:

1. According to (Equ 47) the maximum delay introduced by the egress buffer in Figure 19 is 15μs.

2. According to (Equ 35) the delays introduced by the ingress buffers in Figure 19 are 85μs, 115μs and 85μs.
3. Let the CTD requirement be 100ms for long distance real-time connections represented by the three sources.
4. Let the transmission delay account for 60% of the CTD, i.e., 60ms, as an example.

This gives the total delay (worst case) allowable at the shaper stages for connections using any of the three streams be 39.87ms (i.e., 100ms-60ms-115μs -15μs).

5.6.1. Result of the ISABEL Source

Using (Equ 57) and (Equ 58) and the assumptions that:

- the number of switching stages involved is s=2;
- the maximum delay allowable at the shaper stage is md=39.87ms (as evaluated above);
- mean bit rate is mean=1.8Mbit/s.

the formula for the shaper dimension for the ISABEL source using any of the three QoS streams is (from Equ 57 and Equ 58)

$$\frac{1.8*1000000*y}{424} - 0.03987 shaper\_b = 0 \qquad \text{(Equ 59)}$$

With the QoS based scheme, a set of value for *y* is first determined (based on engineering judgements) at the initial phase of a service implementation. Equ 59 then provides the required buffer sizes for each of the *y* values such that QoS for the service is within a desired boundary; generally, higher the value of *y* selected, the greater is the bandwidth required, leading to better QoS (but higher charges). Which combination of values of *y* and the buffer-size combination are to be used for a given service is then calculated through experimentation in order to take into account service specific requirements. This is because, as was pointed out in section 4.5.5., the exact bandwidth requirement for a connection can be service dependent.

Figure 43 shows the result of the simulations, which input a set of predefined values for *y* and buffer-size.

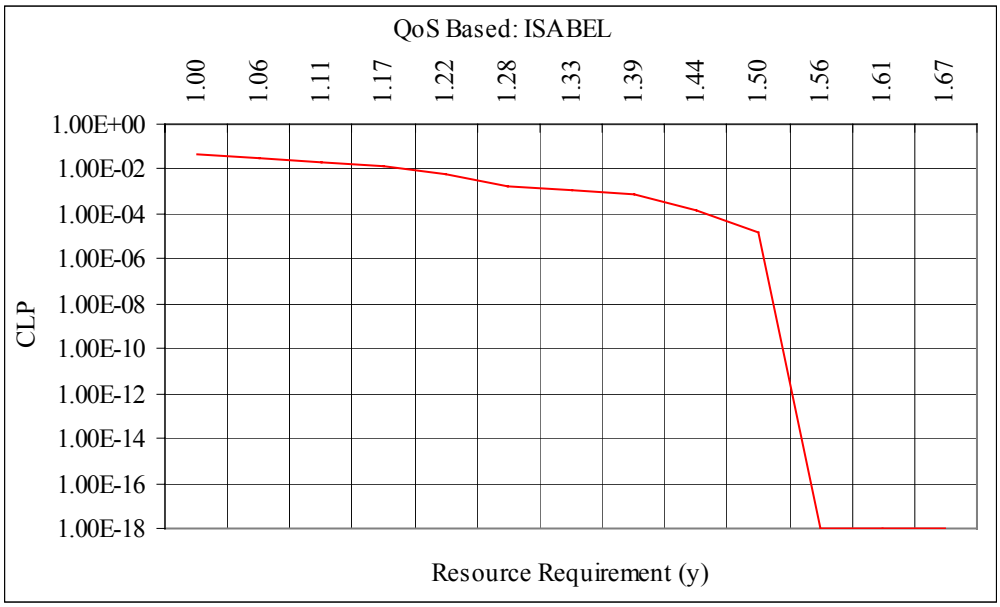

Figure 43: Shaper leak rate, buffer size, and CLP relationship for ISABEL source.

It can be seen from the graph that as y increases, the performance of the shaper improves the traffic stream until the correct value of y is determined for this particular source. At that point, β is considered to be the correct value given the QoS requirement and the correct shaper is in place. For ISABEL, y equal 1.56 and buffer size of 265 (determined through equation Equ 59) is the optimal combination for meeting the QoS requirement and minimising the bandwidth used. *This gives the resource requirement, β, of 2.8Mbit/s.*

5.6.2. Result of the MPEG Source

Using (Equ 57) and (Equ 58) with the assumptions that:

- the number of switching stages involved is s=2;
- the maximum delay allowable at the shaper stage is md=39.87ms;
- mean bit rate is mean=0.6Mbit/s.

the formula for the shaper dimension for the MPEG source using any of the three QoS streams is (from Equ 57 and Equ 58)

$$\frac{0.6*1000000*y}{424} - 0.03987 shaper\_b = 0 \qquad \text{(Equ 60)}$$

With the QoS based scheme, a set of value for y is first determined (based on engineering judgements) at the initial phase of a service implementation. Equ 60 then provides the

required buffer sizes for each of the y values such that QoS for the service is within a desired boundary; generally, the higher the value of y selected, the greater us the bandwidth requirement, leading to better QoS (but higher charges). Which combination of y and buffer-size is to be used for a given service is then calculated through experimentation in order to take into account service specific requirements. This is because, as was pointed out in section 4.5.5., the exact bandwidth requirement for a connection can be service dependent.

Figure 44 shows the result of the simulations, which input a set of predetermined values for y and buffer-size combinations.

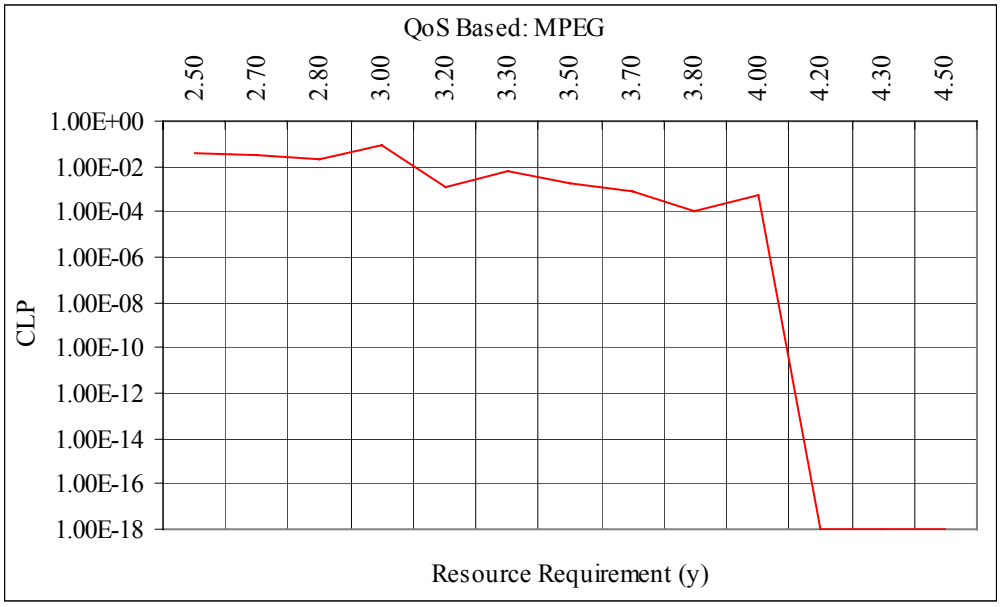

Figure 44: Shaper leak rate, buffer size, and CLP relationship for MPEG source.

It can be seen from the graph that as y increases, the performance of the shaper improves until the correct value of y is determined for this particular source. At that point, β is considered to be the correct value given the QoS requirement and the correct shaper is in place. For MPEG, y equal 4.2 and buffer size of 236 (determined through equation Equ 60) is the optimal combination for meeting the QoS requirement and minimising the bandwidth used. *This gives the resource requirement, β, of 2.5Mbit/s.*

5.6.3. Result of the CBR Source

For the CBR source peak rate of 2.048Mbit/s is taken as the resource allocation and charging parameter.

5.6.4. General Remarks on Simulation Results of the QoS Based Scheme

For the QoS based charging scheme, β has no dependency on the link rate (switch capacity), connection duration or scanning parameters. QoS parameters (CLP and CTD) effect the scheme in two ways:

- The CTD parameter determines the shaper buffer size and the *slr*, which can reduce or increase the size of the resource requirement (β) for individual connections.
- The CLP parameter determines the efficiency of the QoS streams and thereby influences the cost function of all the connections.

Shapers are always necessary for the VBR sources. The customers declare the mean value and the *y* parameter. The mean rate remains stable given a specific service type. On the other hand, customers declare *y* on per connection basis.

It is assumed that when a new service is installed, different values of *y* for varying quality are determined and the mean rate is evaluated. This is performed at the initial phases of the new service. The mean rate together with the range of *y* parameters is then stored in a default configuration (by the user). At connection set-up, customers only need to select the *y* value according to their requirements.

5.7. Comparison and Experimental Evaluation of the Charging Schemes

It has been demonstrated that for the existing charging schemes looked at so far, the resource requirement parameter estimated by the algorithms used by the schemes varies with:

- the link rate supporting the connections;
- the duration of the connections;
- the scanning parameters selected;

all of which would have a negative impact on any resource usage based charging scheme seeking to fulfil the requirements identified in the charging scheme evaluation criteria.

However, in order to reach a final conclusion on the performance of all the charging schemes described in the thesis, an evaluation and comparison process is necessary. Therefore, the author assumes the following model:

- the link rate used to support all the connections is identical;

- the duration of the connection is long enough for all the algorithms to carryout the resource requirement calculations;
- the scanning parameters selected are fair and correct.

The term "correct" refers to the minimum bandwidth reservation judged (through simulations and tests with real network and live traffic) to be necessary in order to receive an acceptable level of pictures and sound quality from the sources under investigation. This was discussed in detail in section 4.5.5.

Given these assumptions, the resource requirements calculated by all the schemes, i.e.:

- $\beta(z)$ for Kelly, (Equ 11);
- d for Lindberger, (Equ 19);
- $k*mean$ for Botvich, (Equ 23);
- eb for Griffiths, (Equ 25);
- β for QoS based charging scheme, together with (Equ 30) and (Equ 31);

are as shown in Figure 45, and Figure 46 for the ISABEL and MPEG sources.

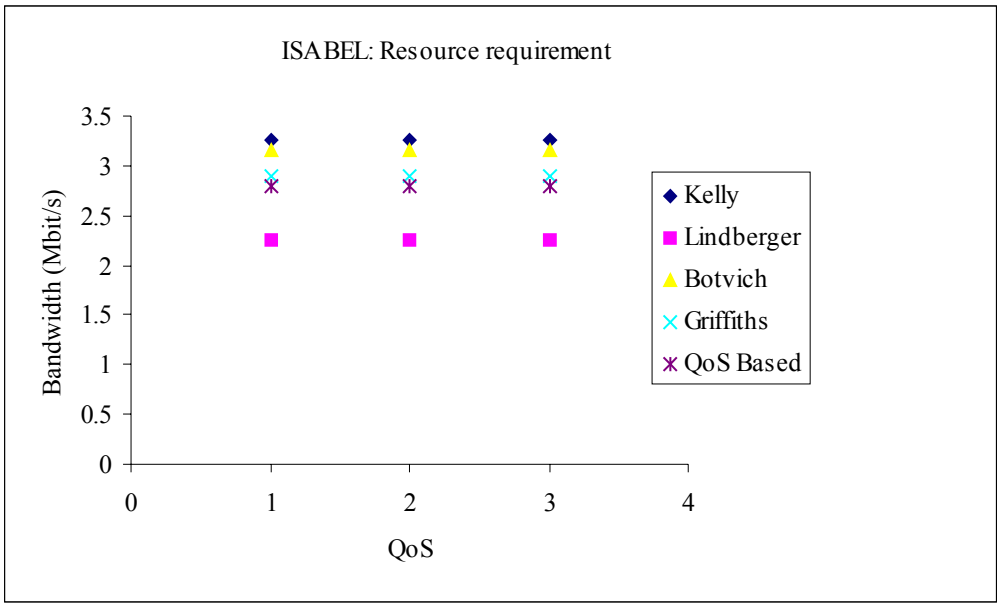

Figure 45: Resource requirement evaluated for the ISABEL source.

For the ISABEL source (Figure 45) notice that the QoS based scheme deduces the resource requirement (bandwidth) of 2.8Mbit/s with traffic shapers. For the rest of the schemes, the estimations are:

- 3.25Mbit/s, Kelly and Botvich;

- 2.25Mbit/s, Lindberger;
- 2.90Mbit/s Griffiths.

Note also that subjective tests with real networks and live traffic have shown (as described in section 4.5.5.) that:

- *the minimum bandwidth required for the ISABEL source (to achieve a reasonable picture and sound quality) is 2.8Mbit/s;*
- *traffic shapers are absolutely necessary if any multiplexing is to be carried out with any other sources.*

Therefore, in this case, the schemes of Kelly and Botvich will heavily overcharge, while the scheme of Lindberger will not calculate a charge correctly (calculation too low).

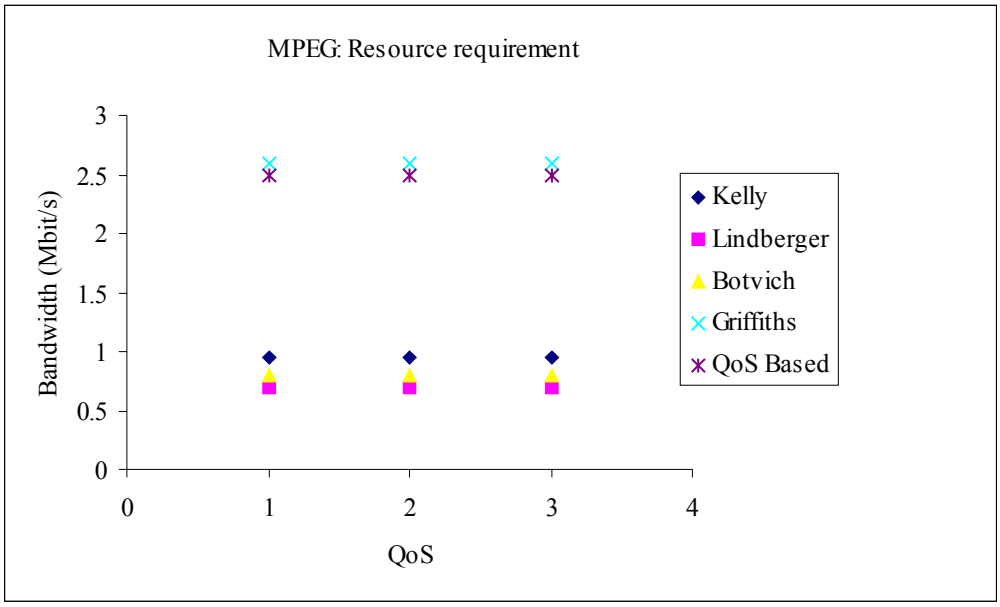

Figure 46: Resource requirement evaluated for the MPEG source.

For the MPEG source (Figure 46), schemes proposed by Kelly, Lindberger and Botvich estimates bandwidth of 0.65-0.95Mbit/s while Griffiths' scheme estimates an equivalent bandwidth of 2.6Mbit/s. The QoS based charging scheme deduces that a bandwidth of 2.5Mbit/s is necessary with a traffic shaper.

Note that subjective tests with real networks and live traffic have shown (as described in section 4.5.5.) that the minimum bandwidth required for this source (to achieve a reasonable picture and sound quality) is 2.5Mbit/s.

This means that, in this case, the schemes of Kelly, Lindberger and Botvich will not calculate the charge correctly (calculated value too low).

The bandwidth used by all the charging schemes for the CBR traffic source is identical, at 2.084Mbit/s.

Having established the resource requirement estimated by each of the charging schemes, and to continue with the comparison work, the author now evaluates real prices for number of connections represented by the traffic sources considered in the simulations.

Assume that the duration of all the connections is 5 minutes. Assume also that market research has shown that at full utilisation (100% capacity being sold) the price of £100 (α in Equ 2) for every 1Mbit/s of bandwidth purchased per minute will recover the set-up cost of a newly installed access network within the first year of its operation. Let this be the target revenue generation for the access network with assumed switch capacity of 155Mbit/s.

Now, with the existing charging schemes, the single buffer shared by all the connections has an efficiency of only 50%, according to Equ 38 (to support the stringent CLP requirement of 10^{-8}). Therefore, the price of the switch usage must be scaled up to £200 (100/0.5) for every 1Mbit/s of bandwidth purchased per minute in order to meet the target revenue generation. This is the price that will be applicable to any connection using resources from the switch.

For the QoS based charging scheme, the access switch is configured to segregate the different QoS streams (QoS of low, medium and high for CLP of 10^{-4}, 10^{-6} and 10^{-8}). The efficiencies of these streams are 85%, 65% and 50%. Therefore, prices applicable to the connections using any of the three streams are £117.65, £153.85 and £200 (α_j in Equ 30).

Using this example, the prices payable by the connections represented by the three sources at varying QoS requirements can be obtained for each of the five charging schemes. Using the resource estimations as: shown in Figure 45 for the ISABEL source; Figure 46 for the MPEG sources; and 2.048Mbit/s for the CBR source, the prices payable are shown in Figure 47, Figure 48 and Figure 49. The x-axis shows the QoS requirement in terms of the CLP and CTD pair. In the above example, the pair is as follows:

- CTD=100ms and CLP=10^{-4} for QoS=Low;
- CTD=100ms and CLP=10^{-6} for QoS=Medium;
- CTD=100ms and CLP=10^{-8} for QoS=High;

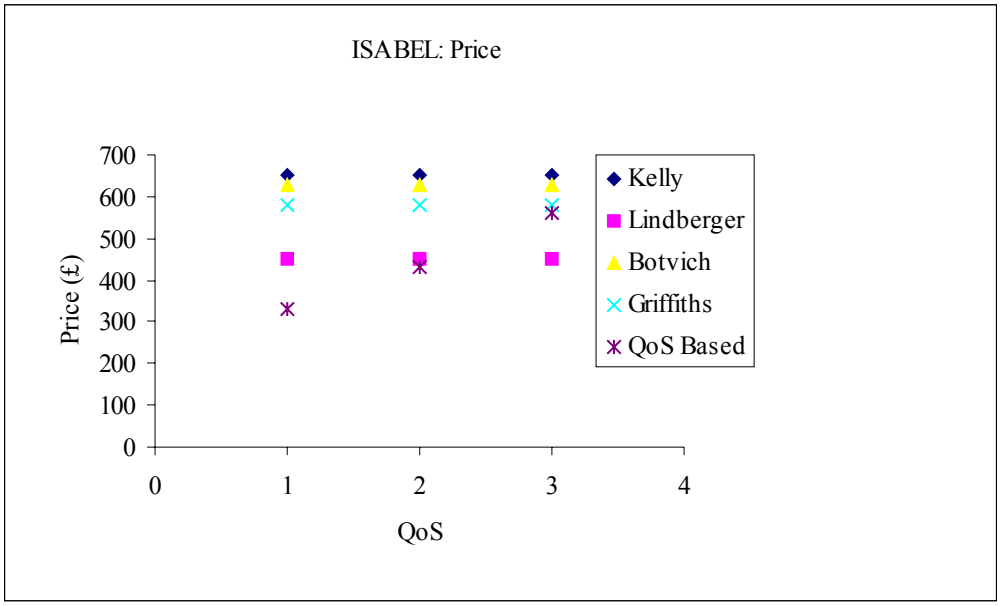

Figure 47: Price evaluated for the ISABEL source.

For the ISABEL source, the price payable through the QoS based charging scheme is comparable with the existing charging schemes (except that of Lindberger) at the highest QoS requirement. However, as the QoS is lowered, the price payable through use of the proposed scheme becomes lower. This happens while the net revenue of the network remains at the target value.

With the scheme of Lindberger, the thresh-hold is at "medium" QoS, above which (at "high" QoS) the price is lower than the QoS based scheme. However, Lindberger underestimates the resource requirement for this source. The consequence of this is that customers will be undercharged, as extra resources will have been allocated by the CAC while charges are based on the estimated bandwidth. This is not satisfactory for the network operator.

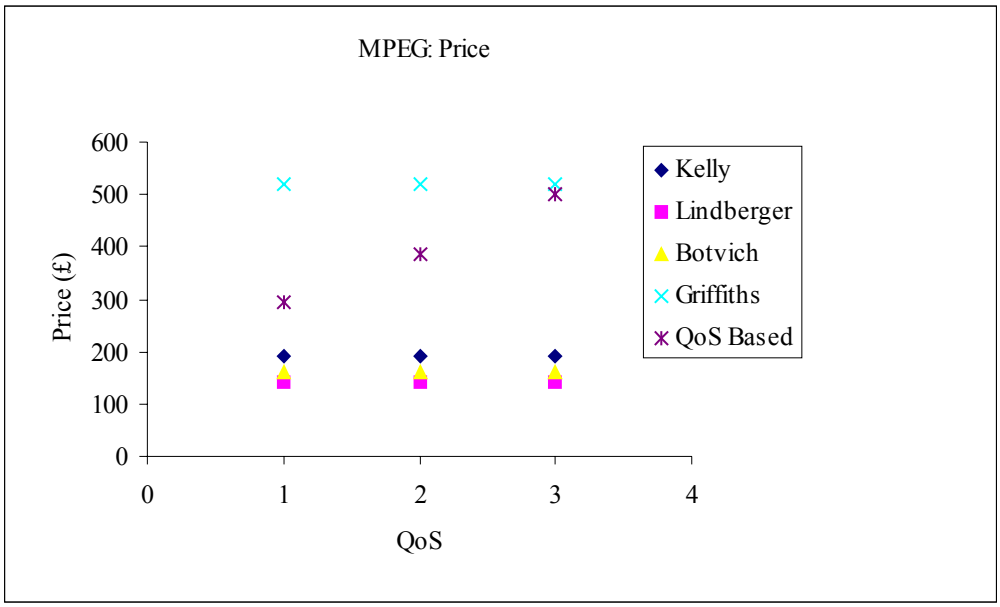

Figure 48: Price evaluated for the MPEG source.

For the MPEG source, similar observations as for the ISABEL source can be observed except that the prices that are calculated by the existing schemes are only comparable with the lower end of the prices deduced by the QoS based charging scheme. This is because the existing charging schemes underestimate the resource requirement for this source: this is not satisfactory for the network operator.

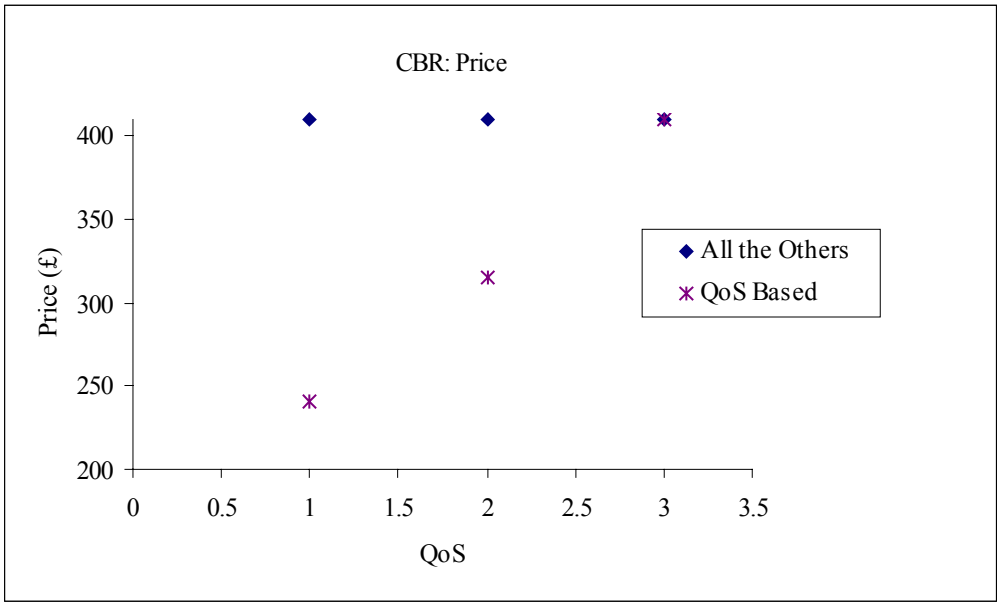

Figure 49: Price evaluated for the CBR source.

For the CBR source, similar observations as ISABEL can be made.

In terms of the charging scheme evaluation criteria, Table 10 summarises the result of the experimental evaluation of the charging schemes. This is the update of the qualitative evaluation process summarised in Table 5, with the addition of the new QoS based charging scheme. Notice that for the existing charging schemes the table remains unaltered. That is, the experiments have confirmed the result of the initial assessment of the charging schemes; that is, the schemes fail on number of criteria. The QoS based charging scheme on the other hand meets all the requirements specified in the criteria list.

	Clear	Account.	Predict.	Flexible	Practical	Control	Choice
Kelly	✗	✓	✗	✗	✗	✗	✗
Lindberger	✗	✓	✗	✗	✗	✗	✗
Botvich	✓	✓	✓	✗	✗	✗	✗
Griffiths	✓	✓	✓	✗	✓	✗	✗
QoS	✓	✓	✓	✓	✓	✓	✓

Table 10: Experimental evaluation of the charging schemes.

5.8. SUMMARY

In this chapter, the performance of all the charging schemes discussed in the thesis was determined through experiments carried out using simulation tools, real networks and live traffic.

It was demonstrated that difficulties could arise with using effective bandwidth as the charging parameter. Also demonstrated was that, if charges are based on the principles that:

- it is the QoS on which customer negotiates the price;
- the resource required remains fixed, dependent only on the service type;

then only the QoS based scheme is satisfactory.

6. DISCUSSION

The main aim of this research has been to:

- Determine the requirements for ATM charging schemes.
- Develop a charging scheme for ATM networks that meets these requirements.

In order to achieve these objectives, initial work focused on identifying all the components that makes up a charging scheme. This resulted in the conclusion, described in sections 3.2., that charges for services run over ATM networks will include:

- service and customer-type related charge;
- subscription charge (SC);
- fixed charge (FC);
- variable charge (VC).

The VC part represents the cost of the actual resource (here bandwidth) used by a connection and it is this resource usage information that has been used in this research. The argument for linking the charging scheme to the measure of resource usage is one of flexibility and fairness: customers pay for what they have used. In a competitive environment, market forces will drive prices closer to the recovery of the cost of the network and this cost must be distributed fairly and in a flexible way to avoid customers taking their business elsewhere.

All the charging components identified above are influenced by two important factors: the type of services being offered and the type of customers using the service. This is part of the market and business strategy of an operator. However, for a charging scheme based on resource usage, the variable charge (part VC) will also be influenced by the customer's perception of the quality of the services they are receiving. Therefore, initial investigation was also required to determine the way in which network service quality are viewed by the customers and how this may be "mapped" onto the charging parameters or tariff parameters that are used in the expression for the variable charges. For that, question and answer sessions were held with representatives from customers and from operators. The result of this discussion led to the proposal of the charging scheme evaluation criteria (CSEC), described in section 3.4. The CSEC was developed to establish a specification list that can be used to design, test and compare ATM charging schemes, and it includes:

1. Predictability: There should be no surprises; charging rate should be known prior to the connection being set-up and shall not change during the lifetime of the connection.
2. Accountability: Capability should be there to retain accurate usage information for auditing purposes.
3. Flexibility: The charging scheme should be service independent, i.e., support all four of the ATM transfer capabilities VBR, CBR, UBR and ABR.
4. Practical: The charging scheme shall not be too complex in terms of implementation; customers need to understand how they are being charged.
5. Control: The charging scheme should allow control of the network in terms of optimising resources and maximising revenue.
6. Choice: The charging scheme should offer choices to the customers in terms of services, quality and cost.
7. Clarity: What services are on offer, how the charges are made and how much is each likely to cost should be clear to the customers.

The criterion identifying "choice" is important. The message that came out of the discussions with the customer group is the importance of choice in QoS, a message that formed the basis for design of the proposed QoS based charging scheme. To appreciate the quality of service of an operator in a competitive environment, customers need to be able to compare their supplier's performance with that of its competitors. Business customers are particularly concerned about quality measures for switched services, dedicated services and virtual services.

Hence, relating QoS to ATM charging schemes is vital if the schemes are to be a success. However, QoS perceived by customers is a subjective measure of quality and therefore it is difficult to quantify QoS in a mathematical equation that forms the expression for the variable charge. To overcome this problem, the author proposes to segregate the QoS requirements between different layers of the protocol stack. The requirements should then be linked to either the tariff parameters, QoS parameters at the ATM layer (CLP, CTD, CDV), or to other requirements existing at a different layer until all the QoS requirements are mapped onto either the tariff parameter or the ATM layer QoS parameters. One mechanism for carrying out this mapping process was presented in section 3.3.

Having established the basis for charging, research concentrated on looking at the existing ATM charging schemes and how they perform in terms of the requirements and expectations (of both the customers and operators) identified in the CSEC. Four of the most widely published charging schemes were selected for detailed analysis (descriptions were presented in section 3.5.).

Having identified the characteristics of the existing charging schemes and a suitable set of criteria, a qualitative evaluation of the schemes was performed as described in section 3.6. Significantly, none of the existing schemes were shown to meet all the criteria specified.

The problems identified with both Botvich's scheme and Griffiths' scheme are that of control, choice and flexibility. Neither schemes allow control in terms of ability to use pricing to better utilise the network resources and maximise profitability. Likewise, customers cannot easily choose different QoS parameters to minimise their network usage cost. Furthermore, the schemes cannot be adopted to support all four of the ATM transfer capabilities (VBR, CBR, UBR and ABR). It can be argued, however, that with Griffiths' scheme customers can choose lower bandwidth to get worse QoS but that is not directly choosing QoS.

With both Kelly's scheme and Lindberger's scheme, there are complications with all but accountability and practicality criteria. Firstly, choice, control and flexibility criteria are not met for the same reasons as for Botvich's scheme. Kelly does use a penalty as the mechanism to force customers to comply with the traffic contract but this is more by the UPC function than using the "control" criterion.

Secondly, both Kelly's scheme and Lindberger's scheme also fail on clarity and predictability. The complex mathematics used to evaluate the charges make it difficult for the customers to understand the basis of the charge – there are no clear charging plans that will always lead to the same charges being made for identical services or connections made at the same time. After the initial declaration of the effective bandwidth requirement of a connection, the charging schemes require on-line measurement of the connection to re-calculate the effective bandwidth for charging purposes. However, the mechanisms used to perform the measurements and estimate the effective bandwidth mean that the final value evaluated is not always predictable. Any small variation in the characteristics of the traffic due to the statistical nature of the sources can result in changes in the effective

bandwidth estimated for a service. This will result in variation of the final charge, leading to surprises that are unacceptable to customers. For example, customers declaring a connection of 2Mbit/s may unexpectedly find themselves at the end of the connection being charged for a 2.5Mbit/s service, without prior warning. In the case of Kelly's scheme this can be significant since the customer will not only pay for the extra 0.5Mbit/s of bandwidth but could also experience cell losses due to the UPC function.

Given the problems associated with the existing charging schemes a new scheme has been proposed (called the QoS based charging scheme, QBCS) which is designed to meet all the criteria identified and which takes into account the importance of choice in QoS. The new scheme was described in detail in chapter 4.

With the QBCS, the customer selects a particular QoS stream depending on the service type (ATM transfer capability) and the QoS requirements. Each stream is pre-configured by the network management system, which takes into account the expected load on the stream, the switch capacity and the expected revenue generation possible from the stream. When a connection is set-up, a cost function is calculated, which will not alter until a new contract is negotiated. The cost is a function of the QoS, the amount of the resource required, plus any other non-resource-related factors the operator may include (such as distance).

The resource requirement is peak cell rate for CBR; shaper leak rate for VBR; volume of cells transmitted for UBR; and volume of cells transmitted and the minimum cell rate requested for ABR traffic.

If the customer underestimates the bandwidth requirement, the UPC will discard cells to protect other customers but no overpayment or penalty will result. If on the other hand they overestimate the bandwidth requirement, improvement in QoS could result and there may be reduction in the charges since the cost is a function of the size of the bandwidth used.

In order to validate the new charging scheme, extensive simulations of a typical network scenario and all the charging schemes were performed. The objective was to test and compare the performance of the schemes experimentally. A common set of services for input was selected to set-up connections, collect measurements, estimate resource requirements and compute charges using a defined price strategy (drawn up for testing

purpose). Full descriptions of simulations and the results obtained from them were provided in chapter 5.

The simulation results confirmed many of the complications associated with the existing charging schemes, as identified in the qualitative evaluation process. Of these, the most significant is the estimation of the bandwidth (effective bandwidth for the Kelly's, Lindberger's, Botvich's schemes and equivalent bandwidth for the Griffiths' scheme). It was demonstrated that in the case of Kelly's, Lindberger's and Botvich's schemes, the estimation is highly unpredictable, being dependent on:

- the link rate (switch capacity) supporting the connections;
- the duration of the connections;
- the scanning parameters used to make the measurements of the connections.

For example, it was observed that customers with identical services but supported by different network interfaces (such as STM-1 and ATM-25) will pay different amount because of the dependency on the link rate.

Furthermore, if a customer does not stay on-line for a sufficiently long period of time (which may actually be quite long – of the order of 10's of minutes), the bandwidth estimation would be unreliable, being too high or too low. Likewise, if the scanning parameters are not set correctly the estimation will also be unreliable; there is no hard-and-fast rule for selecting the correct scanning parameters).

There are additional implementation difficulties that were encountered. Firstly, the effective bandwidth estimated by the algorithms is not always correct in the actual realisation of the services. For example, for the MPEG source used in the experiments the estimated bandwidth of 0.65-0.95Mbit/s is not correct since subjective tests with real networks and live traffic have shown that to get a reasonable picture and sound quality a minimal bandwidth of 2.5Mbit/s is required.

Secondly, in some cases, the bandwidth estimation is correct but a problem occurs in resource allocation. For example, for the ISABEL source the bandwidth estimation is correct (at 2-3Mbit/s). However this cannot be achieved without traffic shapers; if shapers are not used than only a single ISABEL source can be multiplexed into a buffer of service

rate 37Mbit/s (or even higher) without severely degrading the performance of all the connections sharing that buffer.

The difficulties in bandwidth estimation not only introduce uncertainty for the customers but also produce conflicts between the charge and the resource allocation. For example, in the case of Kelly's approach, if the estimation is above that which the customer declared then a severe penalty will be introduced in the final charge while the UPC function might also discard cells. In the case of Lindberger's approach, if the estimation is below that which the customer declared, then charges are based on the estimated value, resulting in extra resources allocated by the CAC for which payment is not received.

Griffiths' method eliminates some of the problems highlighted previously. However, dependency on the link rate does retain many of the uncertainties that have already been highlighted. Furthermore, for certain service types the filter has to be re-dimensioned and this can be time consuming; dimensioning is carried out experimentally using a trial and error approach.

With the QBCS, it has been demonstrated that for a specific service, customers can choose, at any time, any transfer capability and QoS requirement without experiencing any estimation that deviates from the declared parameters. It was further demonstrated that while the resource requirements remained unaltered, the price quoted for a connection varies according to the QoS requested and remains fixed throughout the duration of the connection; there are no differences between what the customer expected to pay and the actual price.

For VBR based services, resource requirements are based on the shaper leak rates. Equations are provided for determining the correct traffic shapers. On implementation of new services, shaper parameters are determined using the equations, and thereafter no alterations are necessary, except for the y parameter (varying form 1 to 5). The y parameter allows fine tuning of the shaper per connection basis, if customers so wish, according to their QoS preferences. QoS streams were created in simulations and consecutive connections were successfully set-up in the simulated network through the streams to demonstrate this capability.

In the case of CBR based services there are no traffic shapers and therefore no *y* parameter to control the resource requirements.

In all cases, it is possible for customers to request too high or too low a bandwidth. If the declaration is too high, extra resources will have been allocated and therefore no reduction in payment is granted although increase in QoS will result. If the declaration is too low then the UPC function will take action to protect other customers but no penalty or extra charges will be made.

With the QBCS the price payable by customers with identical resource allocation but varying QoS requirements is different. This is made possible by distributing the available switch capacity among the QoS streams according to, among other components, the efficiency of the connections supported by each of the stream. That is, resource from low QoS streams (where peak allocation results in overestimation) is distributed to high QoS streams by the network management system. This was demonstrated by setting different loading and service rates of the buffers in the QoS streams according a predefined load distribution factor (normally determined by the network management system).

The overall performance of the QBCS was presented in section 5.7. together with a comparison with the existing charging schemes. By defining a pricing strategy and neglecting any uncertainties with the computations of the effective bandwidth and equivalent bandwidth, the author deduced charges for a number of typical services. This demonstrated the full merit of the provision of choices in QoS selection.

There is however one possible drawback associated with the proposed charging scheme with regard to the total higher layer service independence that the QBCS is designed to have. To support a specific service, customers can run any higher layer applications and choose any combination of transfer capabilities and QoS pairs. However, there are cases, such as Internet services, where a particular combination can cause problems (e.g., [CRO01]). For example, suppose a customer uses TCP/IP protocols and chooses the UBR transfer capability with zero QoS from the network. Because the UBR capability has no inherent quality guarantee, the operator may choose to discard cells from these connections if there is network congestion. This will result in cell retransmissions for lost information. For every ATM cell discarded there will be at least one cell retransmitted (more retransmission will be necessary if the loss of the ATM cell results in discarding of the

entire IP packet). Since the proposed charging scheme assumes volume of cells transmitted as the resource usage information for UBR streams, this will result in an increase in the final charge for added delay but no added benefit to the end service.

The problem highlighted above forms a substantial part of future research (such as those shown in [KIRK01][MAS01]) into higher layer services, service management, and charging. However, before the charging issues can even be discussed there are technical difficulties associated with supporting Internetworking using ATM that must first be overcome. Author of [CRO01] is currently undertaking this challenge.

One further issue remains outstanding from this research, and that is the design of a suitable charging and billing (network) management system to support the proposed charging scheme. A number of requirements for this management system have already been identified in [MIAH02] and [MIAH03] based on [ITU02]. Further requirements are listed in [RCFS90/01].

Finally, it was identified in section 4.5.5. that a scheme such as Kelly's is a "mathematical" effective bandwidth scheme that does not take into account the effect the application has on the resource requirement. That is, bandwidth required can be application dependent rather than just a mathematical manipulation of the traffic profile. The QoS scheme is based on establishing bandwidth to achieve acceptable QoS and in some cases this must be done by subjective means. Further work in this subjective area is required to achieve a full understanding in this area.

At the start of this chapter two main objectives of the research reported in this thesis were stated. In the discussions that followed, the author presented all the steps that were taken to research and develop solutions that will lead to achieving these objectives. The overall theme has been to make ATM marketable to the customers, the network operators and the service providers.

Assisting the success of ATM has been the rational behind the work of this thesis.

7. CONCLUSIONS

Charging for ATM networks is seen to be the final important step in making ATM technology available to customers and service providers. Only through the introduction of a suitable charging mechanism can the bridge between technology and business be made, leading to rapid deployment of ATM.

The charging scheme presented in this thesis is designed to meet the necessary criteria to achieve that acceptability. The techniques used to design and implement the key features of the scheme were evaluated both theoretically and experimentally. In particular, the new QoS charging scheme was tested and compared with number of other existing well-known charging schemes.

A list of criteria was drawn up from the question and answer sessions the author held with representatives from both the customers and operators group. It has been demonstrated that while the existing charging schemes have drawbacks, the proposed scheme meets the necessary requirements of both the customers and the operators by conforming to all the criteria that were identified. In particular the choice of QoS was shown to be the most important factor, and the proposed QoS based charging scheme provides this facility.

During the research two issues that require further study were identified. These were:

- Charging for services supported by the TCP/IP protocols and the UBR transfer capability, with zero QoS support.
- Design of a suitable charging and billing (network) management system architecture to support the proposed charging scheme.
- Further study in the area of "subjective tests" to determine the extent to which bandwidth required by a connection is application dependent.

Of these two issues the first requires substantial research activity and has been identified as such within this thesis.

Overall, the QoS based charging scheme presented by the author allows customers to be charged fairly according to the QoS value they choose. Moreover, it directly relates charge to resource allocation and hence satisfies the requirements of operators.

8. PUBLICATIONS AND REFERENCES

8.1. PUBLICATIONS BY THE AUTHOR

[MIAH01] B. Miah; QoS Based Charging Scheme for ATM Networks; Sixth IFIP Workshop 1998; University of Bradford; Ilkly, UK.

[MIAH02] B Miah; Multi-Provider Multi-Vendor Network & Service Provisioning: Unlocking the Value Chain; The XII International Symposium on Services and Local Access, Venice, Italy, March 1998.

[MIAH03] B. Miah; VPCM OS; Integrated Communications Management of Broadband Networks Edited by D. Griffin; Publisher: Crete University Press, ISBN: 960-524-006-8.

[MIAH04] B Miah; Investigation on Delay and CDV in an ATM-Based Optical Access Network; IEEE ATM97 Workshop, Lisboa, Portugal, May 1997.

8.2. GENERAL REFERENCES

[AAR01] Experimental Investigation of CAC and Effective Bandwidth for Video and Data; E. Aarstad; S. Blaabjerg; F. Cerdan, S. Peters, K. Spaey; 94 ATM Traffic Symposium, Myknos, 1997.

[ATMF01] ATM Forum Traffic Management Specification; Version 4.0; ATM Forum/95-0013R9.

[BAU01] Multi-layer Modelling of a Multimedia Application; M. Baumann, T. Muller, W. Oogh, A. Santos, W. Winstanley, M. Zeller; Broadband Communications 98, Stuttgart, 04/98.

[BEL01] PVC/SVC ATM Networks Design for Voice Traffic; Zheng Chen; Bell Labs; Lucent Technology; Sixth IFIP Workshop 1998; University of Bradford; Ilkly, UK.

[CC01] ATM Charging Schemes; CANCAN Deliverable 5; 1996, AC014/QMW/DS/P/005/B1.

[CC02] Techno-economic Aspects of Contracts and ATM Charging; CANCAN Deliverable 6 (Annexes); 1996, AC014/ANA/DS/P/006/B1.

[COST01] Broadband Network Teletraffic: Final Report of Action COST 242; James Rerts, Ugo Mocci, Jarma Virtamo; Springer 1996.

[COU01] Effective Bandwidth for Stationary Sources; C. Courcoubetis, R. Weber; Prob. Eng. Inf. Sci. 9.

[COU02] Application and Evaluation of Large Deviation Techniques for Traffic Engineering in Broadband Networks; C. Courcoubetis, V. Siris, G. Stamoulis; ACM Sigmetics 98, Performance on Measurement and Modelling of Computer Systems, Madison, Wisconsin, June 24-26, 1998.

[CRO01] Jon Crowcroft, "Why Lossy Internetworking and Lossless ABR ATM Services Do not Go Together"; Research Note RN/94/21.

[CUTH01] L. G. Cuthbert, J. C. Sapnanel; "ATM: The Broadband Telecommunications Solution"; IEE Telecommunications series 29, 1994.

[EXP01] Definition of Optimum Traffic Control Parameters and Results of Trials; EXPERT Deliverable 15; AC094/EXPERT/WP41/DS/R/P/015/B1.

[EXP02] Report on Source Traffic Modelling and Networks Models; EXPERT Deliverable 11; AC094/EXPERT/WP51/D/R/I/011/B1.

[GAB01] Pricing of Telecommunications Services; D. Gabel, M. Kennet; Review of Industrial Organisations, 1993.

[GAR01] A Service Architecture for ATM: from Applications to Scheduling; M. Garrett; IEEE Networks, May/June 1996.

[GRIF01] J. M. Griffiths; Very Selective ATM Statistic Filter; Electronics Letters, Volume 32, No. 7, 28 March 1996.

[GRIF02] J. M. Griffiths; ATM: Customer Needs in the Transition to B-ISDN; Electronics Communications Engineering Journal; October 1996, Volume 8, No. 5.

[GRIF03] Markov Chain Animation, Extension to more than one Dimension and Time Varying Input; 4th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, July 96.

[HEY01] Can ATM Really be this Cheap; P. Heywoods; Data Communications Magazines, Jan 1995.

[HUI01] J. Y. Hui; Resource Allocation for Broadband Networks; IEEE Journal, Selected Areas in Communication, Volume 6, 1998.

[IRV01] The Role of Customer Premises Bandwidth Management; D. Irvin, IEEE Network, May/June 1994.

[ITU01] ITU-T Recommendation I.375 to I.376.

[ITU02] ITU-T Recommendation I.3000 series.

[KELL01] F. Kelly; Notes on Effective Bandwidth of Stationary Sources; Stochastic Networks: theory and applications, Oxford University Press; ISBN 0198523998.

[KIRK01] Business Model and System Architecture for Future QoS Guaranteed Internet Services; P. Kirkby; IEE Colloquium, “Charging for ATM – the reality arrives”, Savoy Place, London, 20 Nov. 1997.

[LIND01] Karl Lindberger; “Cost Based Charging Principles in ATM Networks”; ITC 19, Washington, June 1997.

[MAS01] Pricing the Internet; J. Mackie-Mason, H. Varian; Public Access to the Internet, Prentice Hall, New Jersey 1995.

[MUR01] J. Murphy; Resource Allocation in ATM Networks; PhD Thesis; Dublin City University, March 1996.

[PITTS01] J. M. Pitts, J. A. Schormans; “Introduction to ATM: Design and Performance”; Wiley, 1996; ISBN 0471963402.

[PITTS02] Equivalent Capacity for on/off Sources in ATM; J. Schormans, J. Pitts, K. Willams, L. Cuthbert; Electronics Letters, 13 October 1994 vol. 30, No. 21.

[RCFS90/01] The Accounting and Management Functional Areas; RACE CFS H408; Issue A, CD-NF-13846-EN-C.

[SCH01] Exact Fluid-flow Analysis of a single on/off Source feeding an ATM Buffer; J. Schormans, J. Pitts, L. Cuthbert; Electronics Letters, 1994, 30.(14).pp1116-1117.

[SRI01] A comparative Study of National ISDN versus International ISDN Installations and Tariffs; S. Srinivasan; IEEE Network, May/June 1995.

[STO01] Forecasting long-term Demand for the Services in Residential Market; K. Stordahl, E. Murphy; IEEE Communication Magazine, Feb 1995.

9. APPENDIX A

The information below represents the trial configuration for the QoS charging scheme as shown in Figure 19.

capacity_C:	*155.52M*
ingress_bd1:	*0.25*
ingress_bd2:	*0.33*
ingress_bd3:	*0.25*
ingress_c1:	*3.744M*
ingress_c2:	*4.992M*
ingress_c3:	*3.744M*
ingress_load_1:	*0.80*
ingress_load_1:	*0.65*
ingress_load_1:	*0.50*
egress_b:	*5*
egress_load:	*1.0*
ingress_b1, ingress_b2, ingress_b3:	*10*
shaper_b1, shaper_b2, shaper_b3:	*255 (ISABEL)*
shaper_bj4, shaper_bj5, shaper_bj6:	*265 (MPEG)*
slr1, slr2, slr3 (ISABEL):	*2.8M*
slr1, slr2, slr3 (MPEG):	*2.5M*
pcr (CBR):	*2.028M*
bkg_load_1: 22M; bkg_load_2: 24M; bkg_load_3:	*1.1M*

VCs 1 to 3 are used by the ISABEL sources (3 sources), each traversing through one of the three QoS streams. Likewise VCs 4-6 are used by the MPEG sources while VCs 7-9 by the CBR.

VC-1, VC-2, VC-3:	*ISABEL*
VC-4, VC-5, VC-6:	*MPEG*
VC-7, VC-8, VC-9:	*CBR*
VC-10, VC-11, VC-12:	*BKG Traffic*

The losses occurring at the shapers are listed below. Shapers 1-3 are for the three ISABEL sources, while shapers 4-6 are for the MPEG sources.

CLP_shaper_1:	*0.0*
CLP_shaper_2:	*0.0*
CLP_shaper_3:	*0.0*
CLP_shaper_4:	*0.0*
CLP_shaper_5:	*0.0*
CLP_shaper_6:	*0.0*

Losses at the three streams supporting three QoSs (10^{-4}, 10^{-6} and 10^{-8}) are shown below.

CLP (ingress_1):	*1.0941029879e-04*
CLP (ingress_2):	*1.4343798098e-06*
CLP (ingress_3):	*1.0010000000e-08*

The loss at the egress buffer is shown below.

CLP_egress:	*0.0*

Losses experienced by individual connections (VCs) are shown below. Notice that VCs experienced losses in the order of magnitude supported by each of the streams. For example, VCs 1-3 represents ISABEL sources traversing through the three QoS streams and so the losses are in the order of 10^{-4}, 10^{-6} and 10^{-8}.

CLP_VC1:	*1.0530508470e-04*
CLP_VC2:	*1.8297919996e-06*
CLP_VC3:	*1.0100000000e+09*
CLP_VC4:	*1.2556712336e-04*
CLP_VC5:	*4.1301312021e-06*
CLP_VC6:	*1.0010000000e-08*
CLP_VC7:	*8.7008812425e-05*
CLP_VC8:	*1.0682270040e-06*
CLP_VC9:	*0.0000000000e+00*

The maximum, average and minimum delays experienced by sources regardless of which stream is used are shown below.

MAX_del_(ISABEL/MPEG/CBR):	*280/270/15 cells by 17% of the cells*
AVG_del_(ISABEL/MPEG/CBR):	*175-215/170-200/7-10 by 71% of the cells*
MIN_del_(ISABEL/MPEG/CBR):	*5 cells by 12% of the cells.*

- Note -