Connection Admission Control in ATM Networks

A Thesis
Submitted to the
University of Adelaide
Department of Applied Mathematics
In fulfillment of
The requirement for the degree of
Master of Science (Research)

Eryk Dutkiewicz
Bachelor of Engineering (Honours)
University of Adelaide (1988)

Date of Submission: February 1991

To Clotilde

Preface and Disclaimer

This work has its origin in traffic management and control research studies which I carried out at OTC Limited in Sydney, Australia. It was identified that connection admission control would be one of the crucial aspects required in developing fair and efficient management and control strategies for future ATM networks. OTC Limited has sponsored this research and its outcome is presented in this dissertation.

The work of others is acknowledged and referenced in the text of the thesis. The material contained in this thesis has not been submitted towards another degree or diploma of a university or other institute of higher learning.

The author consents to the thesis being made available for photocopying and loan, if applicable, if accepted for the award of the degree.

Signed

Eryk Dutkiewicz

Acknowledgements

I would like to express my thanks to my supervisors in this work, Dr. Sue Evans (Adelaide University) and Dr. Gary Anido (OTC Limited) for their advice, support and encouragement during this study.

Furthermore, I would like to thank the management of OTC's R&D Section for providing me with the opportunity for carrying out this work while being employed by OTC. A free licence is granted to OTC Limited to use any material in this thesis.

I would also like to thank Mr. David Hughes (Wollongong University) for many useful discussions concerning topics included in this thesis.

Last but not least, I would like to thank my wife Clotilde for her encouragement which helped me complete this work.

Abstract

In this thesis a study of connection admission control for use in future ATM networks is presented. Traffic management and control strategies proposed in the literature are surveyed. An architecture for managing and controlling traffic in ATM networks is then presented. Based on this framework a connection admission model is proposed. Two major aspects of the model, involving the solution of a finite deterministic queue with an arrival process modelled by a two-state Markov-modulated Poisson process and the superposition of two-state Markov-modulated Poisson processes, are presented. The resultant algorithms are then combined to produce an algorithm for connection admission control in ATM networks.

Table of Contents

Chapter One:

Introduction

1.1 Background and Overview of the Thesis

Asynchronous Transfer Mode (ATM) has been identified [9] as a target solution for the implementation of future broadband integrated services digital networks (B-ISDN). A large number of services displaying a wide range of characteristics and demanding from the network distinct and guaranteed levels of performance will have to be supported. The bursty nature of traffic flows will have to be taken advantage of by providing statistical multiplexing in the ATM network. The above requirements indicate a need for developing effective and flexible resource management and traffic control strategies.

The purpose of resource management and traffic control in ATM networks will be to allocate the available network resources in the most efficient and profitable manner in order to support the required services. In order to guarantee distinct performance levels demanded by different services, acceptance of new traffic flows and control of existing traffic flows will need to be carried out.

The problem of accepting new traffic flows into ATM networks has become the centre of research activities in developing resource management and traffic control strategies. It has also received a lot of attention in the standardising activities of the CCITT which is responsible for developing standards for future broadband

networks.

This thesis addresses the problem of connection admission control in ATM networks. The approach taken is, first of all, to investigate methods and techniques for resource management and traffic control proposed in the literature, which involve the problem of connection admission control. As a result of this investigation a flexible general architecture for resource management and traffic control is chosen. Based on this framework, a connection admission control model is proposed and studied. A new connection admission control algorithm is subsequently proposed as a result of the studies.

1.2 Contributions Resulting from Thesis

1. Presentation of a new model for connection admission control in ATM networks.

2. Implementation of a detailed analytical algorithm for the solution of the two-state MMPP/D/1/N queue.

3. Observation of the linear nature of admission regions for heterogeneous traffic streams.

4. Presentation of an algorithm for carrying out connection admission control in ATM network.

5. Presentation of a method for removing individual streams from an aggregate stream without the need for knowledge of the individual characteristics of the remaining streams.

1.3 Publications Based on Thesis

1. E. Dutkiewicz and G.J. Anido "Traffic Management and Control in Broadband Networks", Proc. 5th Australian Fast Packet Switching Workshop, Melbourne, July, 1990.

2. E. Dutkiewicz and G.J. Anido "Connection Admission Control in ATM Networks", Proc. 5th Australian Teletraffic Research Seminar, Melbourne, Dec 1990.

3. E. Dutkiewicz and G.J. Anido "Connection Admission Control in ATM Networks", ITC Specialist Seminar, Cracow, 1991 (to be presented).

1.4 Other Publications

1. E. Dutkiewicz and G.J. Anido, "Traffic Management for Service Quality in Broadband Networks", Proc. 4th Australian Teletraffic Research Seminar, Bond University, Dec. 1989.

2. E. Dutkiewicz and G.J. Anido, "Trunk Controller for use in a Broadband Packet Network", Proc. IREECON International, Melbourne, Sep. 1989.

3. E. Dutkiewicz and G.J. Anido, "A High Speed Trunk Controller for an ATM Switch", Proc. 4th Australian Fast Packet Switching Workshop, Sydney, July, 1989.

Chapter Two:

A Survey of Techniques and Methods for Resource Management and Traffic Control in ATM Networks

2.1 Introduction

The problems of resource management and traffic control in ATM networks have received a lot of attention over the last few years. Problems arise as a result of the requirement for ATM networks to carry a large number of different types of traffic displaying different characteristics and requiring from the network distinct levels of performance. Common network resources are required to support those services and guarantee their performance to users. Resource management functions can be classified into bandwidth allocation and service segregation categories and traffic control functions can be classified into admission control, bandwidth enforcement, and congestion control categories.

The objective of this chapter is to present general methods and techniques which have been identified for resource management and traffic control in ATM networks. The reason for this is that both resource management and traffic control relate to connection admission control. As a result, approaches which have been suggested for resource management and traffic control may also lead to approaches suitable for tackling the problem of connection admission control. Section 2.2 discusses the purpose of resource management in ATM networks and frameworks proposed in the literature which may be used as a basis for addressing problems of providing distinct service performance levels. In Section 2.3 traffic control capabilities, which

have to be provided in ATM networks, are discussed with problems and general solutions identified therein. Since connection admission control forms the topic of the thesis, the problem of connection admission control and models encountered in the literature, which can be used for further study and for algorithm implementation, are surveyed separately in Section 2.4.

2.2 Resource Management

Resource management is required to achieve efficient utilisation of available network resources and satisfy various performance requirements of services to be supported by the network. Two major issues involved in resource management are bandwidth allocation and provision of different qualities of service.

2.2.1 Bandwidth Allocation

In contrast to networks using position multiplexing, where dedicated time slots are allocated for each connection, ATM multiplexing which uses label multiplexing allows virtual bandwidth allocation [55]. Virtual allocation of bandwidth makes it possible for statistical multiplexing to be used in ATM networks. The implication of this scheme is that quality of service can be guaranteed to a particular connection only in a probabilistic sense.

The use of statistical multiplexing for bursty traffic allows more efficient use of available bandwidth to be made when a large number of such traffic streams is combined. Statistical multiplexing, however, increases the complexity of traffic control [62]. On the other hand, nonstatistical multiplexing, where the sum of the peak rates for each traffic stream does not exceed the total available link capacity, provides minimal cell delay and negligible cell loss due to buffer overflow. The disadvantage of nonstatistical multiplexing is the waste of capacity due to low capacity utilisation.

In [29] bandwidth allocation procedures are discussed using a multilayer traffic control and evaluation process. The multilevel traffic process consists of three levels

as depicted in Figure 2.1: packet (cell) level, burst level, and call level. Bandwidth is allocated for each of these levels. At the call level a trunk group is allocated. A trunk group is a collection of transmission links (trunks) bunched together and connecting the same two points in the network. When a burst arrives a particular trunk within the allocated trunk group is selected and individual packets are statistically allocated to available time slots. The multilayered traffic structure with each layer characterised by its own time scale is also proposed in [18] where it is used as a framework for traffic control in ATM networks. An additional 'flow layer' is also introduced in [18] in order to describe variations in the total flow of traffic. Allocation of bandwidth at the burst level is also advocated in [18] and [8]. The disadvantage of such a scheme, however, could be a significant increase in signalling and control traffic. Admission procedures using burst allocation would also require a rapid response time. In addition, for some traffic types burst level description may be meaningless as bursts may not be easily identifiable [8].

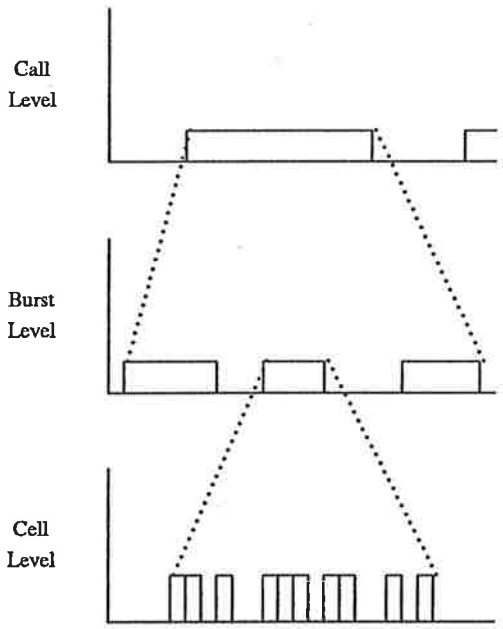


Figure 2.1: Multilevel Traffic Process

2.2.2 Service Segregation

Different services to be supported in ATM networks are expected to demand different qualities of service from the network. If no distinction is made between cells belonging to different connections then the network has to provide a performance level adequate for the most stringent quality of service requirement. This would be equivalent to having only one quality of service for all service types.

Different qualities of service within ATM networks could be provided using mechanisms based on priority schemes. The drawback of these methods is their inflexibility. First of all, only a limited number of priority levels can be provided. This limit is set by the number of priority bits in the ATM cell header. Determining the different priorities in the network might add additional time cost in the processing of cells. Buffer management schemes required for different priorities might also be quite complicated. Although optimisation studies for two priority levels have been carried out [see, for example, 7], no attempt to find the optimal solution for more priority levels has been made. It has also been pointed out [55] that guaranteeing performance levels to traffic with priority levels other than the highest priority might be very difficult. The above problems associated with priority mechanisms suggest that mechanisms based on the concept of virtual paths, as described below, may be more appropriate for provision of guaranteeing qualities of service in the ATM network.

In order to differentiate between different quality of service requirements segregation between services demanding distinct qualities of service could be also performed by partitioning the hardware resources (trunk groups, switch fabrics and so on) into separate subnetworks [55]. This would produce a stringent separation between traffic streams. However, this method might be very inflexible as the number of subnetworks would be limited.

Another method for differentiating between different quality of service requirements is to use an additional layer which would be similar to the virtual path

layer and which would require additional switching and multiplexing facilities [55]. (A virtual path can be regarded as a group of virtual channels sharing common resource through an ATM network [9]). In order to provide such an additional layer, bits in the ATM cell header would have to be allocated for that purpose.

An additional logical layer has been proposed in [2] to associate virtual paths with different qualities of service. A collection of virtual paths providing the same quality of service is termed a virtual network. In order to guarantee such distinct qualities of service, traffic streams belonging to different virtual paths would have to have minimal interference on each other. Any number of virtual networks and hence qualities of service can be provided in the network. However, a large number of such virtual networks might result in low utilisation of transmission resources since no statistical multiplexing of traffic in different virtual paths would be permitted. This point may need further investigation.

2.3 Traffic Control

Traffic control or resource usage control is required to ensure that network bandwidth is allocated fairly between different service classes and that the quality of service for each service class currently using the network is met. The CCITT has proposed four levels of traffic control capabilities which include connection admission control, usage parameter control, priority control, and congestion control. Connection admission control involves functions of the call establishment phase related to a connection. This is discussed further in Section 2.4. Usage parameter control (bandwidth enforcement) deals with traffic monitoring and controlling functions enforcing bandwidth allocated to connections. Priority control makes use of the cell loss priority bit as recommended by CCITT [9] and provides a mechanism for accommodating different priority traffic flows in the network. The primary use of this mechanism is expected to be made in congestion control. The aim of congestion control is to react to and minimise the effect of traffic overload.

Preventive control is advocated by most people as it may not be feasible to carry out flow control at the cell level due to the high speed of operation of ATM networks. A general framework, in the context of congestion control in ATM networks, has been proposed in [61]. It advocates preventive control with the use of access control in order to achieve fair congestion control in an efficient manner. Congestion controls based on access control can be broken down into three levels: route control, admission control, and bandwidth enforcement. Route control is used to ensure efficient resource utilisation in the network, whereas admission control and bandwidth enforcement are used to ensure a required performance level to all users. Multiple performance classes are also proposed in [61] as a means of achieving bandwidth efficiency while guaranteeing performance levels required by services. In one approach, available bandwidth would be segregated for each class, requiring physical separation of network resources. (This is a similar approach to that suggested in [55] as discussed earlier). Another approach suggested in [61] is the use of explicit priority levels. However, as was discussed earlier, priority levels may be too restrictive.

The authors also point out in [61] the need for standard traffic descriptors which would be used for traffic flow characterisation in admission control and bandwidth enforcement. However, it is not clear which characteristics in a traffic stream should be used in order to accurately determine its queueing performance. This is still an open question which is being addressed [43], [53].

Such traffic descriptors could also be used in the derivation of effective bandwidth values. An effective bandwidth is a number associated with a traffic stream and it describes bandwidth requirement of the stream taking into account the characteristics of the stream, the characteristics of other streams sharing the same transmission resources, the available transmission capacity, and the performance level required by the traffic stream. The concept of the effective bandwidth is discussed further in Section 2.4.

Admission control can be subdivided into two categories: call admission control and connection admission control [55]. Call admission control functions determine the availability of destination subscriber, availability of network facilities to support the type of call requested, charging aspects of the call and other features related to negotiation for network resources. A call can consist of a number of connections, as in case of a multimedia or conference call. Individual connections belonging to the call must be coordinated during the call establishment phase. Connection admission control is responsible for determining whether a connection requesting a particular quality of service should be accepted or rejected. The decision should be based on the characteristics of the new traffic flow to be admitted and the current state of the resources. Note also that the distinction between call admission and connection admission is not made in a lot of papers.

The characteristics of the traffic flow for which the connection is requested should be used in the connection admission algorithm and should correspond to traffic flow parameters which are used by the bandwidth enforcement method. Section 2.4 discusses models for connection admission control proposed in the literature.

2.3.1 Bandwidth Enforcement

Bandwidth enforcement or traffic enforcement is required to ensure that traffic flows conform to the agreed bandwidth usage. Its purpose is to protect the network resources from malicious or unintentional violation of traffic volume agreed during the connection phase. This guarantees that qualities of service of other connections are protected.

Bandwidth enforcement performed at the access points to the network is also known as source policing [4]. Other functions which could be performed by source policing units at such points include security functions [9], collection of statistics, and charging information.

The possible use of traffic shapers in conjunction with traffic enforcers has been also widely suggested [61], [4], [9], [16]. Traffic shapers could be used for traffic flow conditioning and they would differ from traffic enforcers in that they would buffer violating cells rather than discard or tag them. A possible configuration in which traffic shapers and enforcers could be used is shown in Figure 2.2.

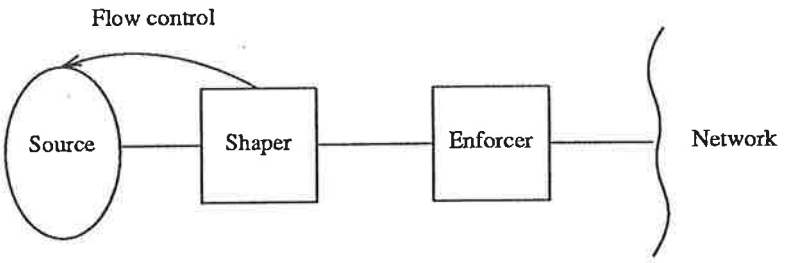


Figure 2.2: Possible Configuration for Preventive Bandwidth Enforcement with Use of Traffic Shapers

Traffic shaping could be carried out either at the user end or it could placed together with the traffic enforcer at the network interface.

Bandwidth enforcement mechanisms can carry out the enforcement by discarding cells which violate the agreed traffic volume parameters at the access point to the network. The other method suggested in [62] and [16] proposes tagging violating cells as they enter the network and then treating them as having the lowest priority ensuring preferential discard under network overload conditions. Both methods have been included in the current CCITT standards.

A study of different bandwidth enforcement mechanisms has been presented in [48]. Four mechanisms which are studied include the leaky bucket, the jumping window, the moving window, and the exponentially weighted moving average mechanism. The principle of operation of the leaky bucket as well as the principle of its hardware implementation is depicted in Figure 2.3. It consists of a counter which is incremented when a cell arrives and decremented at a constant rate. The counter has a predefined limit which can be reached after the arrival rate exceeds

the decrement rate. When the limit is reached any subsequent cells are discarded until the counter value falls below the limit. The leaky bucket can be modelled by a G/D/1/K queue where K corresponds to the counter limit and the service rate of the queue corresponds to the decrement rate. The important distinction between the leaky bucket and a real queue is the lack of storage and hence no delay suffered by cells in the leaky bucket.

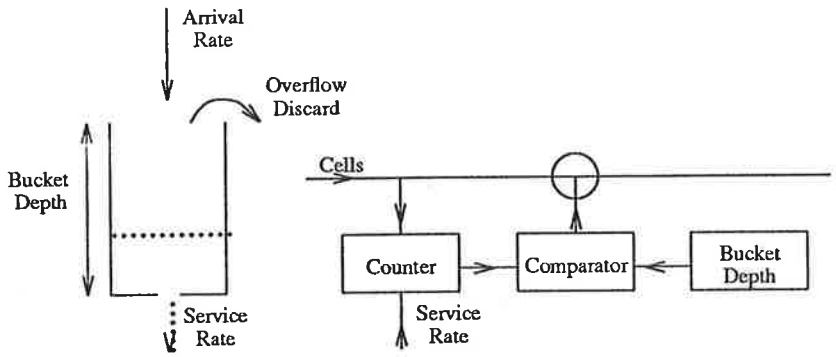


Figure 2.3: Principle of Leaky Bucket Operation and Hardware Implementation

In the jumping window mechanism the number of cells permitted within a fixed time interval (window) is limited to a maximum number N. The next window starts immediately at the end of the previous window. The counter keeping track of the number of cells is restarted at the beginning of each window. As a result, a cell can influence the value of the counter during the time interval ranging from zero to that corresponding to the size of the window.

The moving window mechanism is similar to the jumping window mechanism. In this case, however, the maximum number of cells within a given interval T is limited in any possible interval of length T. Thus instead of the window 'jumping', as is the case in the jumping window mechanism, in this case the window is steadily moving along the time axis. The moving window mechanism can be modelled by a G/D/N/N queue, where the number of servers N corresponds to the maximum number of cells allowed within an interval T and deterministic service times of the

servers are given by 1/T.

The exponentially weighted moving average (EWMA) mechanism is also similar to the jumping window mechanism except that the maximum number of cells N_i permitted in the i-th window is a function of the allowed mean number of cells per interval N and an exponentially weighted sum of the number of cells S_{i-1} which were accepted in previous windows such that:

$$N_i = \frac{N - \alpha S_{i-1}}{1 - \alpha} \text{ where } 0 \leq \alpha \leq 1$$

The effect of the past in the mechanism is determined by the value of α. When $\alpha = 0$ the past has no effect on the system. (i.e. the system becomes equivalent to the jumping window mechanism)

The comparison of the above mechanisms carried out in [48], which takes into account the cell loss probability of the source, sensitivity to static overload, dynamic reaction time, and the worst case traffic admitted by each mechanism, concludes that the leaky bucket and the EWMA are the most promising mechanisms (of the four examined) for use in bandwidth enforcement. In the moving window mechanism arrival times of cells must be remembered and to achieve the same cell loss probability as other mechanisms large counter values are required. This makes the implementation of this mechanism expensive.

A lot of attention has been given in the literature to the problem of traffic characterisation which can be used in bandwidth enforcement. Traffic descriptors used for such characterisation must be easy to estimate and enforce. However, they must also allow for characterisation of a wide range of traffic flows. Le Boudec [35] suggests negotiation directly in terms of leaky bucket parameters. These would have important implications on traffic shaping by users. The idea of negotiating traffic controller parameters can be extended to controlling mechanisms other than the leaky bucket. Such approach, however, might make the network too implementation-specific.

In [33] traffic sources are characterised by mean and variance of the bit rate distribution which is approximated by a Gaussian distribution. The amount of bandwidth allocated to connections is based on these approximations. For connections departing from a Guassian distribution more information on the amount of deviation from the distribution is required, resulting in weighted mean and variance parameters used in the bandwidth allocation formula. As far as admitting new connections is concerned, the performance measure is the probability that the cell rate for all multiplexed connections do not exceed a predefined maximum value.

Bandwidth enforcement is proposed in [33] to be carried out by means of load counters. They are used to monitor the difference between the distribution of the cell rate of the connection in progress and the Gaussian approximation defined by call parameters presented during the call set-up. A Gaussian distribution can be approximated using a 'staircase' method as shown in Figure 2.4. However, bandwidth enforcement based on departures from the Gaussian distribution seems restrictive, especially as far as individual traffic streams are concerned. It may be difficult to justify Gaussian approximations for a large number of traffic types expected in future ATM networks. Such approximations also ignore time correlation and burstiness in traffic streams.

The effectiveness of traffic bandwidth enforcement functions is questioned in [22] and [12]. The difficulties in specifying source characteristics with accuracy and implementing enforcement mechanisms, which will not adversely affect well behaved sources but which will penalise sources which do not adhere to their 'contract', are the main reasons for concern. In particular, in [12] the effectiveness of the leaky bucket mechanism for mean bit rate policing is studied. The performance of this mechanism has been found to depend critically on the distributions of the on and off periods for an on/off source and not just on their mean value. It is suggested that the solution to the above problems will be obtained in the future by a combination of different forms of traffic control acting on

different types of traffic.

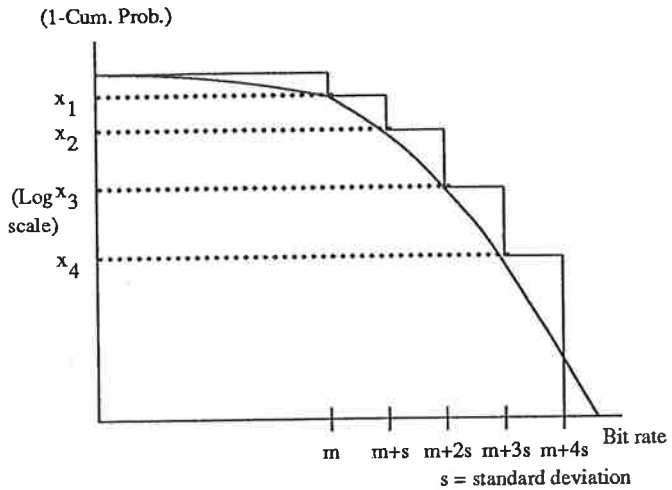


Figure 2.4: Staircase Approximation for Gaussian Distribution

2.3.2 Congestion Control

Congestion control, which is treated here as a function of traffic control, involves actions carried out by the network in order to protect network resources from overload conditions. The purpose of such actions is to minimise congestion effects and to avoid the spread of congestion [9]. Traditional methods of flow control used for congestion control may not be used in ATM networks due to high transmission rates which would require rapid reaction time from controllers. Rather than using reactive control mechanism preventive control schemes have been suggested. Preventive control is particularly suited to the connection oriented mode of operation of ATM networks. These methods aim to ensure that network traffic intensities never reach congestion levels, thus treating congestion as a rare event due to unpredictable statistical fluctuations of traffic flows [9] or failure in the network. Preventive congestion control can be carried out using connection admission control and bandwidth enforcement control. While in this manner long term congestion (such as caused by a network failure) may be minimised, short term congestion (such as caused by statistical fluctuations in traffic streams) still needs to be addressed.

Although preventive control has been suggested most commonly, cell discard mechanisms using some form of reactive control have also been proposed. Main work in this area has been carried out by Eckberg, Lucantoni and associates. A framework for congestion, flow, and error control based on bandwidth management is proposed in [13]. Congestion control is implemented within the network, whereas flow and error controls are implemented in end devices. Traffic entering the network is monitored in real time and cells which exceed agreed traffic parameters are marked with a violation tag and are not guaranteed to be transported. Control devices distributed along the flow path discard violation tagged cells when congestion has to be relieved. This mechanism may allow for better utilisation of the transmission resources compared to discarding all violation tagged cells at the entry points to the network. However, the impact of congestion, due to tagged cells being allowed into the network, on the non-tagged cells in the network has not been determined yet.

The leaky-bucket algorithm which is used for traffic monitoring in real time is treated in [16]. Negotiation at call set-up is necessary to determine leaky-bucket parameters for traffic monitoring process. The network should be able to determine whether the requested traffic parameters can be supported. The purpose of this study has been to characterise the output of non-tagged traffic taking into account the effect of burstiness on the relevant time scale of the network. The study is carried out using a two-state Markov modulated Poisson model (which is described in detail in Chapter 4) for the source traffic and dynamics of the leaky bucket are modelled by a finite capacity queue with a deterministic service. In order to study burstiness of non-tagged cells in terms of the time scale of the network, the non-tagged traffic, which passes through the leaky bucket, is subsequently put into an infinite-server system with exponentially distributed service times (with rate parameter μ). The parameter μ is used to capture the effects of network utilisation. The conclusion drawn by presenting a numerical example is that the impact of bursty traffic on network performance is greatest when burst dynamics of traffic

entering the network are short compared to network time-constants.

The technique for assessing burstiness in terms of time scales relevant in the network described above is based on the concept of peakedness. Peakedness, as used for characterisation of teletraffic processes, is described in more detail in [15]. This type of characterisation allows approximations of interactions of a traffic stream with a range of service systems. Peakedness is defined as:

$$z(\mu) = \frac{var[X(\infty)]}{E[X(\infty)]} \tag{2.1}$$

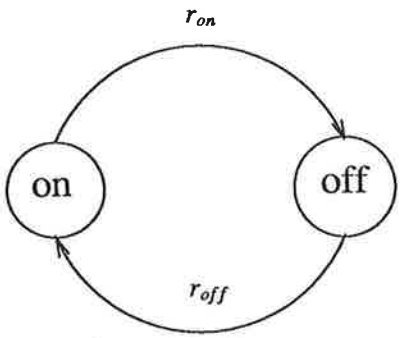
where $X(\infty)$ denotes the number of customers in the infinite-server system. Service times of the server are exponentially distributed. This definition can be generalised [15] to allow service times to have any distribution with a finite mean. Peakedness provides a quantitative method for assessing effects on congestion produced by various characteristics in both the arrival stream and the server. Congestion is caused by the interaction between the arrival stream and the queueing system as a result of variability in the arrival stream over the duration of the time that an arrival spends in the system. The concept of peakedness will be also used in Chapter 5 in reference to traffic superposition.

2.4 Connection Admission Control Models

Models for connection admission are necessary in order to determine conditions which will allow efficient utilisation of resources in ATM networks where different types of calls will need to be supported. The two components needed in such models are the traffic model and the ATM connection model. Approaches taken in the literature for addressing the problem of connection admission control can be broadly categorised as queueing model approaches and approaches which are based on the concept of the effective bandwidth. These approaches are discussed below.

A queueing model for connection admission control has been proposed in [32]. A source model based on the on/off model is used to represent traffic flows. When the

source is in the on-state it transmits cells with a fixed rate and when it is in the off-state it is silent. Durations of the on and off states are exponentially distributed. Figure 2.5 shows a transition diagram for the on/off traffic model.



Average time in on-state = $\frac{1}{r_{on}}$

Average time in off-state = $\frac{1}{r_{off}}$

Figure 2.5: Model for On and Off Traffic Source

Cell loss probability is chosen as a measure of quality of service required by services. A single link modelled by a cell multiplexer with a FIFO queue is used for quality of service evaluation. It is assumed that the most restrictive quality of service requirement has to apply to all services. The cell loss probability of 10^{-9} corresponding to this quality of service is chosen.

Evaluation of the link performance is carried out using four queueing models with traffic input in all cases modelled as on/off sources. The first model is a *fluid flow queueing model* for heterogeneous traffic and it is treated as the reference model for the remaining models. The second model (called the *homogeneous fluid flow approximation*) is produced as an approximation of the reference model in which the arrival process is approximated by identical on/off sources. By approximating the queueing system in the second model by a corresponding multi server loss

system the third model is produced. This model is referred to as the *binomial multi-server approximation* since the number of sources, which is at any time in the on-state, is binomially distributed. This is equivalent to the model proposed in [20] which was analysed using a large deviation approximation. (Study presented in [20] is also discussed later in this section). The last model is constructed in a similar fashion by approximating the queueing system in the reference model (which used heterogeneous on/off traffic) by a multi-server loss system. Since the input stream into this system can be approximated by an arrival process having a negative binomial or Pascal distribution, this model is referred to as the *pascal multi-server approximation.*

The four models are compared using numerical calculations. It is concluded that fluid flow models are too complex for obtaining admission control criterion and an admission control algorithm based on the *binomial multi-server approximation* (i.e. multi-server loss approximation to the homogeneous fluid flow approximation) is advocated. This is an interesting result as traditionally fluid flow approximations produce simpler solutions than those obtained from queueing models. It could be attributed to the fact that the simplification achieved in the *binomial multi-server approximation* sacrifices information regarding the distribution of the on and off periods in the traffic source.

Another approach to connection admission is presented in [59]. Traffic generated by a connection is described by stochastic processes Z_n, Y_n, and by a random variable X which correspond to cell level, burst level, and connection level descriptions.

Cell loss probability of the ATM link used by connections is related to the permissible throughput of the link and depends on the architecture of the ATM switching element. This cell loss probability is taken as an input parameter describing the performance of the link. This parameter could be obtained by analytical or simulation methods at the cell level. Cell loss probability of an

individual connection is then obtained in terms of cell loss probability of the link and traffic characteristics on the link.. It is noted that fixed cell loss probability in an ATM link does not give the same cell loss probability for each connection utilising the pipe. This observation, however, is made by considering different periods of time for connections displaying different characteristics.

The resulting admission algorithm uses the mean and peak bit rates of the new and existing connections to decide whether a new connection should be accepted. The algorithm also requires that all connections in the same link are segregated into two classes. Connections belonging to one class are allocated peak bit rates whereas those belonging to the second class are suitable for statistical multiplexing. However, it is not made clear how interference between the two classes is to be taken into account.

In [20] connection admission problems are studied assuming a bufferless ATM network. Traffic sources are assumed to alternate between periods of activity and inactivity. During periods of activity sources are transmitting at a fixed bit rate, whereas during periods of inactivity sources are silent (cf the model in [32] discussed earlier). They are defined using two parameters: peak bandwidth when active, and the fraction of time corresponding to the active time. The statistical behaviour of the system is studied using a large deviation approximation. The criterion for connection acceptance is a prespecified cell loss probability allowed in a transmission resource. Cell loss probability is defined as the probability that the instantaneous demand of active sources exceeds the available capacity of the resource. An example is given in [20] in which three different types of connections are combined. One of the interesting results obtained in this study is the near planarity of the surface defining the permissible admission region for these connections. Such surfaces are found to be almost planar for a wide range of parameters. It is suggested that this near planar nature of connection admission boundaries could be used as the basis for a simple connection admission control strategy in which each traffic type would be assigned an effective bandwidth. The

effective bandwidth of a source would depend on the characteristics of the source, capacity of the transmission resource and the allowed cell loss probability. It would, however, be independent of characteristics of other sources. A new connection would be accepted if the resulting sum of the effective bandwidths of all the connections in progress does not exceed the available transmission capacity. Further questions relating to the definition of a "traffic type" and the number of different traffic types allowed using this strategy may still have to be investigated.

Other strategies for connection admission control based on the idea of the effective bandwidth have also been proposed. In [40] a survey of such methods is presented. A general requirement for cell level performance is represented as:

$$f_i^W(n_1, ..., n_k) \leq P_i \quad for \quad 1 \leq i \leq k$$

where f_i^W represents a performance function, n_i the number of connections of class i, P_i the cell level performance requirement for the class i connection, and W the available transmission capacity. The above inequality specifies the admissible region for connection acceptance.

Methods for determining above regions which use effective bandwidth are called indirect methods as opposed to direct methods in which the calculation of the performance function f_i^W is carried out on-line for each new state $(n_1, ..., n_k)$. A number of possible approximations are identified using indirect methods. In the simplest method the effective bandwidth $W(n_1, ..., n_k)$ for all connections sharing a transmission link is given as the sum of the peak rates of each connection [61], [17]. This method, however, is very inefficient as it does not take advantage of statistical multiplexing at the cell level. In [17] more efficient use of available capacity is made by assuming that capacity is overallocated. This is done in order to take into account random fluctuations in the bit rate requirement of each connection.

Another approximation for calculation of the effective bandwidth is given as:

$$W(n_1, ..., n_k) = \sum_{i=1}^{k} W_i(n_i)$$

where $W_i(n_i)$ represents the effective bandwidth for class i and it is defined as the minimum value satisfying:

$$f_i^Z(n_i) \leq \min P_j$$

where $f_i^Z(n_i)$ represents cell level performance when only class i connections are carried by the transmission link with capacity Z. This approximation takes advantage of statistical multiplexing of connections within each class, however, statistical multiplexing between different classes is ignored.

Another approximation for obtaining effective bandwidth which is reviewed in [40] is given by:

$$W(n_1, ..., n_k) = \sum_{i=1}^{k} n_i V_i$$

where V_i is the effective bandwidth for a class i connection, and is given by link capacity W divided by n_i^{max}. n_i^{max} is the maximum number of class i connections which can be carried on the link so that the cell level performance requirement is met, i.e.

$$f_i^W(n_i^{max}) \leq P_i.$$

This approximation is also suggested in [20]. It is, however, pointed out that the above approximation is too optimistic and methods for improving its accuracy are more complex [40]. One of the methods which attempts to achieve more accuracy is proposed in [14]. The evaluation of the effective bandwidth in [14] takes into account interaction of different classes of traffic. However, the method is state dependent and requires more complex calculations.

Effective bandwidth approximation using Gaussian distribution have also been proposed. They are, however, not suitable for bursty traffic and methods which attempt to improve its accuracy are also more complex [33], [58].

2.5 Conclusion

This chapter has presented main approaches and techniques found in the literature for resource management and traffic control in ATM networks. The overall aim of such approaches has been to satisfy performance requirements of diverse services expected in future ATM networks and at the same time to make efficient use of available network resources.

While individual aspects of this problem have been addressed, an overall framework is required which will allow flexible provision of multiple performance levels and efficient utilisation of network resources. One of the objectives of Chapter 3 is to present a new traffic management and control architecture based on the concept of the virtual network.

Particular attention has also been given in the literature to the problem of connection admission control, where a number of different models have been suggested. The architecture presented in Chapter 3 will be used as a framework for studying connection admission control in ATM networks. A queueing model for connection admission based on that framework will be also proposed in Chapter 3.

Chapter Three:

Traffic Management and Control Aspects in ATM Networks Leading to Connection Admission Control

3.1 Introduction

Quality of service has been used as a performance measure describing the level of performance experienced by a service in the network. Different definitions of quality of service used in different types of networks stem from the fact that those networks were designed to support one particular service and hence their performance was optimised for that service only.

In ATM networks a large number of distinct services will have to vie for the same network resources. Different management and control schemes will have to be devised in order to support diverse quality of service requirements of such services. One aspect of traffic management and control which will be affected by different quality of service requirements is connection admission control.

In this section quality of service as a performance measure in future ATM networks is discussed. This in turn leads to control requirements necessary to provide distinct qualities of service in ATM networks, taking into account the statistical nature of such networks. A traffic management and control architecture is presented next and using this framework a connection admission control model is proposed. The connection admission control model consists of two parts: the queueing model for

the virtual path and the arrival stream model for stream superposition in a virtual path. Analytical studies of these two parts are carried out in Chapter 4 and 5 respectively.

3.2 Control and Service Quality Considerations

Future ATM networks will have to provide 'bandwidth on demand' to network users. They will have to handle various types of traffic requiring different bit rates and demanding different qualities of service. Unlike in traditional circuit switching networks, bandwidth in ATM networks will not be fixed to a limited number of allowed levels. In ATM networks users will be able to demand bandwidth from a continuous and wide spectrum. This has important implications for managing and controlling the use of network resources.

The term quality of service has been used to cover a range of performance requirements which users impose on networks. The concept of quality of service of telecommunications services has been a major focus of the standardising activities of the CCITT. It has defined the term for telephone networks and narrowband ISDN networks in its recommendations. According to its definition the overall quality of service as seen by users depends on a number of factors. It is characterised by the combined effect of service support performance, service operability performance, and service integrity as defined in [9]. The service integrity factor concerns transmission performance and more precisely the level of reproduction of the transmitted signal at the receiving end [9].

The two main parameters which specify quality of service in ATM networks are end-to-end delay and cell loss probability. These parameters become particularly relevant when network resources are statistically assigned to services. When network resources are allocated deterministically (i.e. according to peak bit rate requirements), cell delay and cell loss are minimal and approach those of circuit switching networks [44]. In ATM networks, where high speed trunks will be used, cell delay may be negligible and cell delay control may not be required. This leaves

cell loss probability as a main performance measure or quality of service measure and control of this parameter will be necessary if statistical multiplexing is to be used. However, such measures as maximum cell loss variation, and probability and duration of periods of high cell loss rates may also have to be considered [62]. Nevertheless, only cell loss probability will be regarded as an important performance measure in this study.

Another performance measure which may be important in ATM networks is grade of service. Grade of service can be considered in terms of connection blocking probability in contrast to quality of service which is measured in terms of cell blocking probability when the connection is already in progress. In this study only quality of service will be considered, however, provision for inclusion of grade of service in the connection admission algorithm will be made.

The requirement for "bandwidth on demand" and for multiple qualities of service to be provided by ATM networks has important implications on the control capabilities which the network has to display. Bandwidth allocated to users will have to be controlled to ensure efficient use of network resources. Quality of service provided in the network will also have to be controlled. This can be achieved by controlling individual traffic streams and by ensuring that their effect on each other results in the required performance level.

ATM networks can be envisaged as providing virtual transport networks. These transport networks allow flexible use of available bandwidth and due to the statistical nature of such networks it is possible to achieve multiplexing gain with traffic displaying bursty characteristics. In contrast to the statistical nature of the transport network, control of resources and control of quality service will have to be carried out in a deterministic fashion.

Connection admission control can be regarded as a traffic control function as outlined in Chapter 2. It concerns admission of new traffic streams into the network based on the current state of the network and the performance level to be satisfied.

The admission control model, which will be described under the architectural framework, will be based on the concept of preventive congestion control as described in Chapter 2.

The above considerations lead to the need for a flexible traffic management and traffic control architecture which will take advantage of the statistical nature of ATM networks and which will provide deterministic control capabilities including admission control capabilities.

3.3 Traffic Management and Control Architecture

The architecture presented here was first proposed in [2]. While this architecture is not the result of the studies presented in this thesis, it is used as a framework under which a connection admission control model is proposed. The architecture consists of three levels: ATM Resource Management level, ATM Traffic Management level, and ATM Traffic Control level. These levels relate to resource management, admission control, and bandwidth enforcement which were discussed in Chapter 2. The architecture is based on the concept of virtual networks.

The concept of virtual networks forms an important part of the traffic management and control architecture. A virtual network (or a logical overlay network) is defined as a set of virtual paths created in the network in order to provide a given minimum quality of service to network users. A number of virtual networks may be set up depending on the number of distinct qualities of service to be supported. A connection using a particular virtual network is guaranteed the quality of service provided by the virtual network. During the lifetime of the connection, however, the actual quality of service experienced by the connection may in fact be better than the minimum guaranteed in the virtual network. Figure 3.1 depicts a representation of virtual networks created to satisfy a number of quality of service requirements

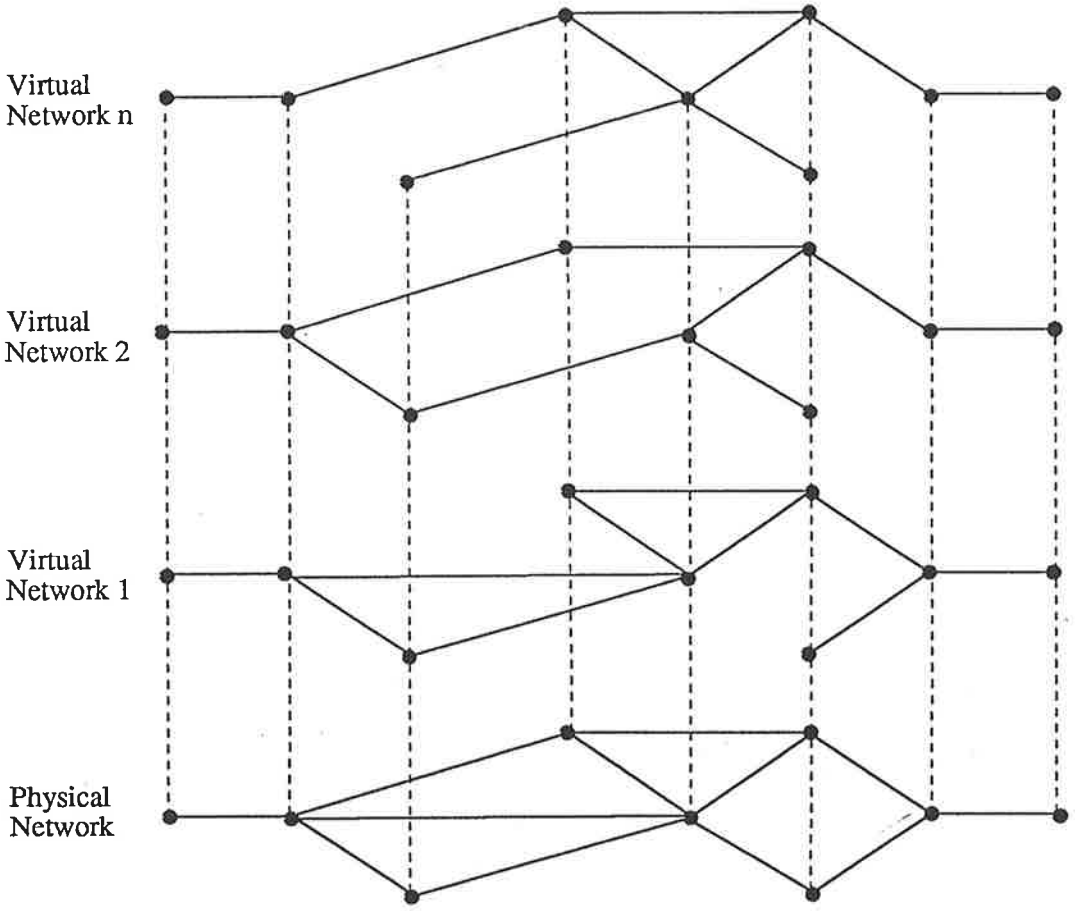


Figure 3.1: Virtual Networks Based on Quality of Service Requirements

Each of the levels forming the traffic management and traffic control architecture is described and discussed below. The architecture is also depicted in Figure 3.2 showing the three levels.

3.3.1 ATM Resource Management

The ATM Resource Management level is responsible for setting up and managing virtual networks as demand for them changes in the network. This level monitors the use of resources in virtual paths and adjusts the allocated resources as demand falls or rises. It receives requests from ATM Traffic Management units for more capacity and it reduces available capacity if capacity is being underutilised. Failures in the network will also affect resource allocation action at this level.

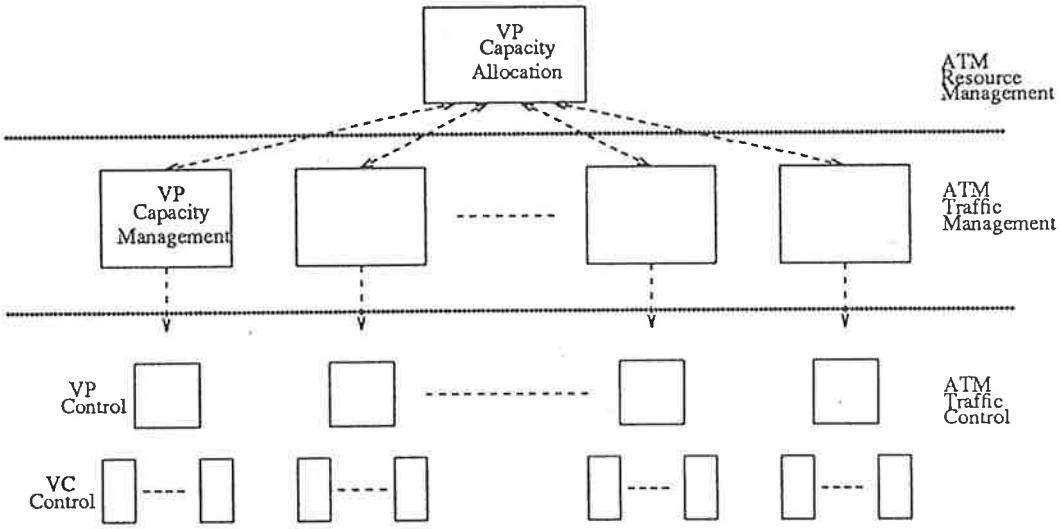


Figure 3.2: Traffic Management and Control Architecture

3.3.2 ATM Traffic Management

The ATM Traffic Management level is responsible for managing capacity in individual virtual paths. At this level decisions are also made whether a connection should be accepted into the network. Admission decisions will have to be based on traffic parameters describing traffic flows in connections, the quality of service and the grade of service demanded by connections, and the resources available to the network. Traffic classes may be defined in the network based on the limited number of possible traffic characteristics which the network can recognise. Every connection will then be assigned a traffic class which closely matches its characteristics. Similarly, a limited number of virtual networks can be provided corresponding to the qualities of service supported by the network. Grade of service classes may also be defined based on the grade of service required by connections.

3.3.3 ATM Traffic Control

The ATM Traffic Control level is responsible for ensuring that traffic flows in individual virtual channels and virtual paths do not exceed allowed thresholds. As discussed in Chapter 2, due to the high speed of information flow in ATM networks,

preventive control as opposed to reactive control will have to be exercised. The main purpose of this usage monitoring mechanism at the virtual channel level is to protect network resources from users who by exceeding their negotiated traffic parameters might otherwise degrade the quality of service of other users. The simplest, and most conservative, action that can be taken upon detecting traffic violation is cell discard in the offending stream.

At the virtual path level the aim of the traffic control level is to maintain separation between different virtual paths so that any interference between traffic belonging to those virtual paths is minimised. This action will ensure that qualities of service provided in virtual paths belonging to different virtual networks are protected. A simple way of implementing this separation can be achieved by policing each virtual path according to the peak bit rate corresponding to the capacity of the virtual path. Figure 3.3 depicts a representation of virtual channels and virtual paths in an ATM link.

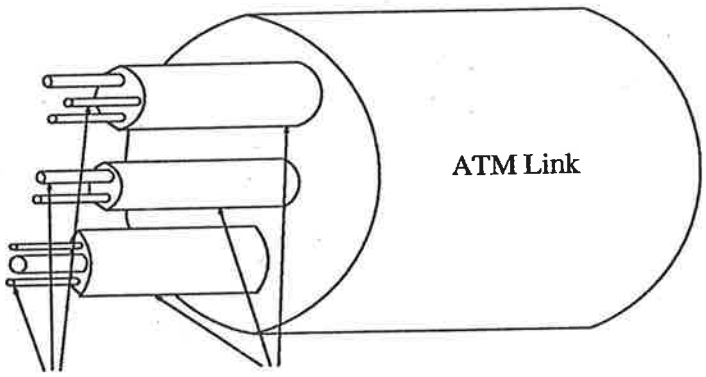


Figure 3.3: Virtual Paths and Virtual Channels in ATM Networks

Note that different time scales are associated with each level in the architecture. ATM Resource Management is expected to carry out its operation within the period of tens of minutes or longer. ATM Traffic Management which deals with new connections will have to respond according to the connection arrival and termination rate resulting in response time in the order of milliseconds. Lastly,

ATM Traffic Control will have to handle individual cells which will have interarrival times of microseconds. These time scales have implications on the interactions between the different levels and the models which can be applied at each level.

3.4 Connection Admission Control Model

In the simplest model for connection admission control, a connection will be admitted into the network if the network can guarantee the quality of service required for the new connection and the connections already in progress. Certain traffic statistics of the connection will have to be presented to the network to allow the decision to accept or reject the connection to be made. The admission mechanism will have to determine if there exists a virtual path into which the particular connection can be assigned and if there is enough spare capacity for that connection.

Each connection requires a virtual channel to be set up between the required origin and destination, which will provide and guarantee the required quality of service. The virtual channel will be part of the virtual path, which can be envisaged as carrying other virtual channels which also require similar quality of service from the network. Depending on the type of traffic in the virtual channels, some statistical multiplexing may be allowed (or even desirable if traffic is of bursty nature). Each virtual channel in a particular virtual path will experience the same performance from the network as the other virtual channels sharing with it the virtual path. In order to ensure that the quality of service of each individual virtual channel is maintained, each virtual channel must be guarded so that any excess traffic, differing from the agreed traffic statistics, is not allowed to enter the virtual path. All virtual paths are separated from each other in order to ensure independence of traffic streams carried in those virtual paths. This ensures that traffic streams belonging to different virtual paths do not interfere with each other thus guaranteeing their qualities of service. Hence, to ensure that traffic streams in

different virtual paths are protected from each other, each virtual path should be guarded.

The guarding or policing mechanism, both at virtual channels and virtual paths will have to be inexpensive (so that additional cost per customer is minimal), but at the same time it will have to be effective. As discussed in Chapter 2, a leaky bucket mechanism has been proposed as a way for policing real time traffic. In particular, the policing mechanism at the origin of each virtual path should ensure separation of traffic belonging to different virtual paths and hence qualities of service provided by those virtual paths. In order to achieve this separation traffic policing based on peak bit rates will have to be enforced.

Based on the above observations and using the traffic management and traffic control framework introduced earlier, a model for connection admission control can be proposed.

The connection admission model can be separated into two parts: a queueing model for the virtual path and an arrival stream model for stream superposition in a given virtual path. The criterion for admission of a connection is the overall cell loss probability experienced by connections in a given virtual path.

3.4.1 Queueing Model for Virtual Path

At the virtual path level a traffic stream representing the aggregate of all traffic streams carried in the virtual path is considered. The quality of service of all connections in that virtual path is then determined by the cell loss probability set by the policing mechanism for the virtual path. As pointed out earlier, since separation of traffic in different virtual paths must be ensured in order to provide and guarantee distinct qualities of service, the policing mechanism will limit the peak bandwidth allowed in the virtual path.

The size of buffers in ATM networks is expected to be small. Large buffers might increase cell delay and cell delay jitter of delay sensitive service [33]. In this model

of connection admission control a measure of quality of service is taken as the overall cell loss probability suffered by traffic streams in the network. Taking into account the expected cell loss probability of less than 10^{-9} for future ATM networks, it is assumed that cell loss will be imposed on traffic streams predominantly by policing mechanisms. In fact, at these entry points to the network service dependent loss probability could be introduced with ease using the above mechanisms.

Taking the above requirements into account, an admission mechanism employing a policing mechanism based on peak rate limiting can be represented by a short queue with a deterministic server (see Figure 3.4). As a result, the model reduces to that of a G/D/1/N queue, where G denotes a general arrival process, D a deterministic server, and N the number of waiting spaces (including the server) in the queue. It should be noted that this queue affects only the cell loss probability of the entering stream. No cells are actually queued and the cell stream entering the remainder of the network is not the output process of the above queueing system. To achieve complete separation between virtual paths, the number of waiting spaces in the queue should be set to $N = 1$. However, to allow some bursts to take place in the arrival stream at the peak bit rate and to obtain better utilisation of the available bandwidth, the queue length might be set somewhat higher. The question of setting the queue length (or the depth of the leaky bucket) will be investigated in Chapter 5.

A two-state Markov-modulated Poisson process can be used as the arrival process into the above queue. The two-state MMPP will be chosen in such a way as to approximate the aggregate stream in the virtual path. The two-state MMPP process is described in detail in Chapter 4.

3.4.2 Stream Arrival Model

The connection admission model also involves the superposition of a number of two-state MMPPs representing traffic streams in individual virtual channels. A two-state MMPP is a very general arrival process which is analytically tractable and

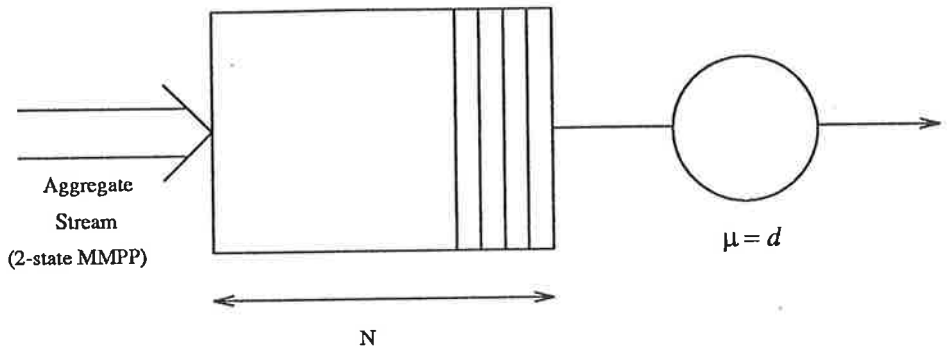


Figure 3.4: Connection Admission Queueing Model

which has been used to model a wide range of traffic sources [38], [52]. Traffic streams corresponding to virtual channels need to be superimposed in order to obtain an aggregate traffic stream which can be applied to the queueing system modelling the policing mechanism of the virtual path as described in the previous section. Figure 3.5 depicts the required stream superposition.

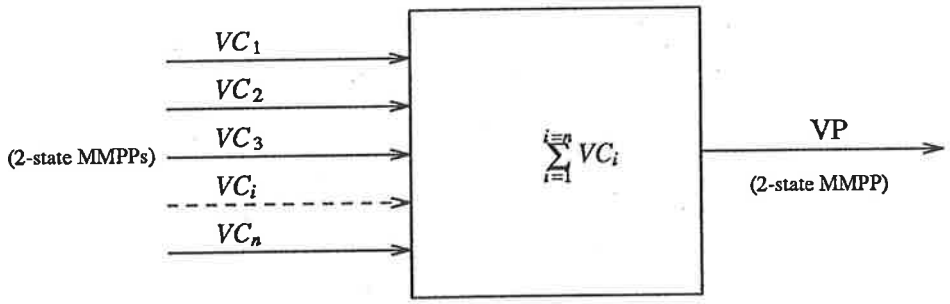


Figure 3.5: Stream Superposition in Connection Admission Model

A traffic stream in a virtual channel must pass through a traffic controller placed at the origin of the virtual channel. This traffic controller imposes an upper limit on the bit rate of the stream. Action of traffic controllers, however, should not change the characteristics of the traffic stream drastically, unless the stream is violating its negotiated parameters. On the other hand, if violation does occur then the stream which is output by the traffic controller implemented using a leaky bucket has been found to have less effect on cell loss probability of other traffic streams in the

network, compared to its effect when negotiated parameters are adhered to [27]. As a result, a traffic stream which is modelled by a two-state MMPP according to the user-specified characteristics and which is being policed according to those characteristics should result in the worst case scenario as far as its effect on the cell loss probability experienced by other traffic present in the virtual path is concerned [28].

Although traffic in individual virtual channels is proposed to be represented by two-state MMPP in this model, other descriptions of such traffic may also be considered. Superposition methods suitable for obtaining the aggregate stream from individual traffic streams will be required. In Chapter 5 a superposition method, which does not depend on the assumption that traffic in individual virtual channels is modelled by a two-state MMPP, will be studied. However, due to the versatility and simplicity of the two-state MMPP description, this description will be maintained in the connection admission algorithm in Chapter 6.

The method for stream superposition must be able to capture the important characteristics of the aggregate stream and it must be suitable for implementation in a connection admission algorithm. Two issues which are important, as far as implementation is concerned, are addition of a new connection to the existing aggregate and removal of an old connection from the aggregate.

3.5 Conclusion

A flexible traffic management and traffic control architecture has been presented. Using this architecture efficient use of available network resources can be made. It can also provide distinct and guaranteed quality of service levels. The architecture consists of three levels at which resource management, admission control, and bandwidth enforcement take place respectively.

A connection admission model has been proposed using this architectural framework. Two distinct parts of the model have been identified. These are the

virtual path queueing model and the stream arrival model for superposition of traffic streams in individual virtual channels belonging to the same virtual path. The queueing system, modelled by the two-state MMPP/D/1/N queue is analysed and solved for cell loss probability (representing quality of service) in Chapter 4. Traffic superposition of two-state MMPPs modelling traffic in individual virtual channels is subsequently studied in Chapter 5. The criterion for connection admission is the overall cell loss probability suffered by all the connections in a given virtual path. The connection admission control algorithm based on the above connection admission model is proposed in Chapter 6.

Chapter Four:

Analytical Techniques Leading to Algorithmic Solution of Virtual Path Queueing Model

4.1 Introduction

The proceeding chapter presented an architectural framework for traffic management and control. Based on this framework a connection admission control model was proposed. The model was separated into two parts. The first part contained the model of the virtual path in which the virtual path was represented by the two-state MMPP/D/1/N queue. The arrival model for this queueing system, representing the superposition of individual connections in a virtual path, constituted the second part of the overall model. The objective of this chapter is to analyse the first part of the connection admission control model and to provide a solution for the two-state MMPP/D/1/N queue in terms of cell loss probability, which is regarded as the measure of quality of service provided in a virtual path.

In Section 4.2 a Batch Markovian Arrival Process, of which the two-state MMPP is a special case, is studied. Section 4.3 presents an analysis of the BMAP/D/1/N queue. This analysis is used in Section 4.4 to obtain an algorithmic solution of the two-state MMPP/D/1/N queue.

4.2 General Arrival Processes Leading to two-state MMPP

A two-state Markov-modulated Poisson process has been widely proposed as a traffic model for a large variety of traffic types ranging from data to voice and video [16], [38], [46], [52], [53], [30], [34]. One important characteristic of the two-state MMPP is that it can be used in obtaining analytically tractable solutions when used as an input to queueing systems.

A generating model for the two-state Markov-modulated Poisson process is shown in Figure 4.1. The process can be seen to result from an alternated switching between two Poisson processes. These processes are characterised by the intensities λ_1 and λ_2 respectively. The times that the process spends in each of the two states are exponentially distributed with means T_1 and T_2. Hence the two-state MMPP can be completely characterised by the following four parameters:

$$\lambda_1, \lambda_2, r_1 = \frac{1}{T_1}, \text{ and } r_2 = \frac{1}{T_2}.$$

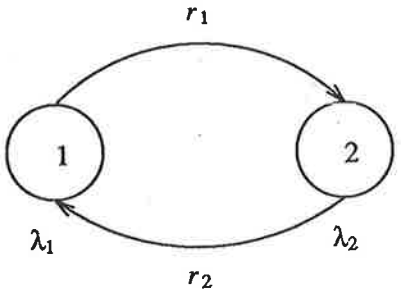


Figure 4.1: Two-state MMPP

A two-state MMPP can be regarded as a special case of two more general processes presented in [38] and [42]. These processes, known as the Batch Markovian Arrival Process (BMAP) and the Versatile Markovian Point Process (VMPP) respectively, are equivalent to each other. The notation introduced for the BMAP is, however, simpler and more elegant. The analysis of these general processes leads to a

particularly simple description of the two-state MMPP, which can be used in the analysis of the two-state MMPP/D/1/N queueing system. The subsection below presents the main features of the BMAP with this aim in mind.

4.2.1 Batch Markovian Arrival Process

Consider a 2-dimensional Markov process {N(t), J(t)} on the state space $\{(i,j): i \geq 0,\ 1 \leq j \leq m\}$ with **Q**, the infinitesimal generator given by:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \mathbf{D}_4 & \cdots \\ 0 & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \cdots \\ 0 & 0 & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots \\ 0 & 0 & 0 & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ 0 & 0 & 0 & 0 & \mathbf{D}_0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix} \tag{4.1}$$

where $\mathbf{D}_k$, $k \geq 0$ are $m \times m$ matrices.

If N(t) represents a counting variable and J(t) a phase variable then the above process defines a batch arrival process where transitions from a state (i,j) to state $(i+v,l)$, $v \geq 1$, $1 \leq j,l \leq m$ correspond to a batch arrival of size v. The evolution of the process can be described as follows. Let the Markov process describing the process of moving between the phases (called from now on the underlying Markov process) be in phase j, $1 \leq j \leq m$. The time during which this process will remain in this phase is exponentially distributed with parameter λ_j. At the end of that time a transition occurs to another or back to the same phase. This transition may or may not correspond to an arrival epoch. Transitions to the same phase without an arrival are, however, not considered. With probability $p_j(0,k)$, $1 \leq j,k \leq m$, $k \neq j$, there will be a transition to phase k without an arrival. With probability $p_j(v,k)$, $v \geq 1$, $1 \leq j,k \leq m$, there will be a transition to phase k with a batch arrival of size v. Thus:

$(D_0)_{jj} = -\lambda_j$,

$(D_0)_{jk} = \lambda_j p_j(0,k)$, $1 \le j, k \le m$, $k \ne j$,

$(D_v)_{jk} = \lambda_j p_j(v,k)$, $v \ge 1$, $1 \le j, k \le m$.

Thus matrix $\mathbf{D}_0$ governs transitions which correspond to no arrivals, and $\mathbf{D}_v$ governs transitions which correspond to arrivals of batches of size v. Figure 4.2 depicts the evolution of the Markov process with generator $\mathbf{Q}$.

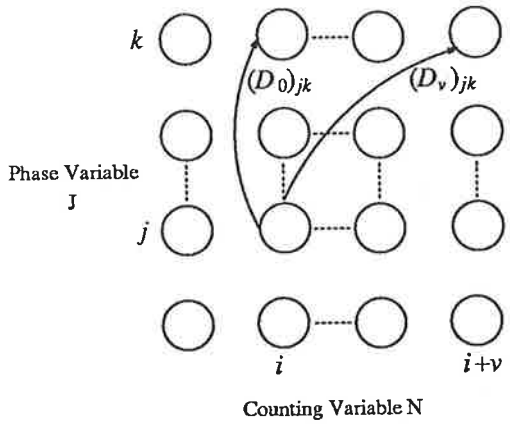


Figure 4.2: Evolution of the Markov process with generator $\mathbf{Q}$

Let $\mathbf{P}(t)$ be the transition probability matrix of the Markov process {N(t), J(t)}, with generator $\mathbf{Q}$. Then the forward Chapman-Kolmogorov equation for the Markov chain becomes:

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}, \text{ for } t \ge 0, \text{ with } \mathbf{P}(0) = \mathbf{I}.$$

Let

$$P_{jk}(n,t) = P\{N(t) = n, J(t) = k \mid N(0) = 0, J(0) = j\} \tag{4.2}$$

be the (j,k) element of an $m \times m$ matrix $\mathbf{P}(n,t)$. Matrix $\mathbf{P}(n,t)$ can be obtained from the transition probability matrix as the n-th block matrix in the first row of $\mathbf{P}(t)$.

The transition probability matrix has then the following form:

$$\mathbf{P}(t) = \begin{bmatrix} \mathbf{P}(0,t) & \mathbf{P}(1,t) & \mathbf{P}(2,t) & \mathbf{P}(3,t) & \mathbf{P}(4,t) & \cdot & \cdot \\ 0 & \mathbf{P}(0,t) & \mathbf{P}(1,t) & \mathbf{P}(2,t) & \mathbf{P}(3,t) & \cdot & \cdot \\ 0 & 0 & \mathbf{P}(0,t) & \mathbf{P}(1,t) & \mathbf{P}(2,t) & \cdot & \cdot \\ 0 & 0 & 0 & \mathbf{P}(0,t) & \mathbf{P}(1,t) & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \mathbf{P}(0,t) & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

The matrices $\mathbf{P}(n,t)$ satisfy:

$$\mathbf{P}'(n,t) = \sum_{v=0}^{n} \mathbf{P}(v,t)\mathbf{D}_{n-v}, \quad n \geq 1, t \geq 0, \tag{4.3a}$$

$$\mathbf{P}(0,0) = \mathbf{I}. \tag{4.3b}$$

The matrix generating function $\mathbf{P}^*(z,t)$ can be defined as:

$$\mathbf{P}^*(z,t) = \sum_{n=0}^{\infty} \mathbf{P}(n,t)z^n, \quad \text{for } |z| \leq 1. \tag{4.4}$$

It satisfies:

$$\frac{d}{dt}\mathbf{P}^*(z,t) = \mathbf{P}^*(z,t)\mathbf{D}(z),$$

$$\mathbf{P}^*(z,0) = \mathbf{I},$$

and can be obtained from:

$$\mathbf{P}^*(z,t) = e^{\mathbf{D}(z)t}, \quad \text{for } |z| \leq 1, \ t \geq 0. \tag{4.5}$$

where $\mathbf{D}(z)$ is the matrix generating function defined as:

$$\mathbf{D}(z) = \sum_{i=0}^{\infty} \mathbf{D}_i z^i, \quad \text{for } |z| \leq 1 \tag{4.6}$$

Also define $\mathbf{D}$ as:

$$\mathbf{D} = \sum_{k=0}^{\infty} \mathbf{D}_k. \tag{4.7}$$

and let π be the stationary probability of the Markov process with generator **D**, so that π satisfies:

$$\pi \mathbf{D} = 0,$$

and

$$\pi e = 1.$$

Generator **D** describes the underlying Markov process.

The Markov-modulated Poisson process (MMPP) is a special case of the batch Markov arrival process (BMAP) described above. The MMPP has an infinitesimal generator **R** for the underlying Markov process given by:

$$\mathbf{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & \cdots & r_{2,m} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & \cdots & r_{3,m} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & \cdots & r_{4,m} \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdots & \cdot \\ r_{m,1} & r_{m,2} & r_{m,3} & r_{m,4} & \cdots & r_{m,m} \end{bmatrix}$$

and an arrival rate matrix Λ given by:

$$\Lambda = diag\,(\lambda_1, .., \lambda_m).$$

In this case all arrivals consist of a batch of size 1 and hence $\mathbf{D}_j = 0$, $j \geq 2$. Also $\mathbf{D}_0 =$ R - Λ, and $\mathbf{D}_1 = \Lambda$. Arrivals cannot occur at the same time as the transitions between states of the underlying Markov process, hence the form of $\mathbf{D}_1$.

For a two-state Markov-modulated Poisson process we have:

$$\mathbf{R} = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}$$

and Λ is given by:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Hence $\mathbf{D}_0$ and $\mathbf{D}_1$ are given by:

$$\mathbf{D}_0 = \begin{bmatrix} -r_1 - \lambda_1 & r_1 \\ r_2 & -r_2 - \lambda_2 \end{bmatrix}$$

and

$$\mathbf{D}_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

and hence the infinitesimal generator $\mathbf{Q}$ is given by:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & 0 & 0 & 0 & \cdots \\ 0 & \mathbf{D}_0 & \mathbf{D}_1 & 0 & 0 & \cdots \\ 0 & 0 & \mathbf{D}_0 & \mathbf{D}_1 & 0 & \cdots \\ 0 & 0 & 0 & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ 0 & 0 & 0 & 0 & \mathbf{D}_0 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdots \end{bmatrix}$$

with $\mathbf{D}_0$ and $\mathbf{D}_1$ as above. Note that for a MMPP, $\mathbf{D}$ is equivalent to $\mathbf{R}$. Figure 4.3 shows the evolution of the two-state MMPP.

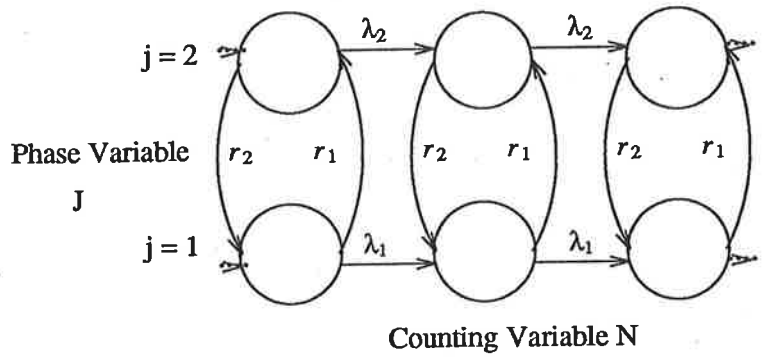


Figure 4.3: Evolution of the two-state MMPP

4.3 Analysis of the BMAP/D/1/N Queue

In the following section the matrix-analytic approach to the BMAP/G/1/∞ queue will be presented. This will be followed by the analysis of the finite version of the above queue - the BMAP/G/1/N queue. The analyses presented below are based on the works of [38], [42], [46], and [6].

4.3.1 Analysis of the BMAP/G/1/∞ Queue

Define the arrival process by the sequence $\{ \mathbf{D}_k , k \geq 0\}$. Then the fundamental arrival rate for the process is given by:

$$\lambda = t^{-1} \pi \frac{d}{dz} \mathbf{P}^* (z,t) \mathbf{e} \tag{4.8}$$

evaluated at $z = 1$, where π is the stationary probability vector of the underlying Markov process with the generator $\mathbf{D} = \sum_{k=0}^{\infty} \mathbf{D}_k$, and $\mathbf{P}^*(z,t)$ is the generating function for the transition probabilities $\mathbf{P}(n,t)$ as defined in Equation 4.4. Also λ^{-1} is the fundamental (or average) interarrival time of the process. The above expression can be simplified further by noting that:

$$\frac{d}{dz} \mathbf{P}^* (z,t) = t \frac{d}{dz} \mathbf{D}(z) \, e^{\mathbf{D}(z)t}$$

Then from Equation 4.6:

$$\frac{d}{dz} \mathbf{D}(z) = \sum_{k=1}^{\infty} k \mathbf{D}_k z^{k-1}$$

$$= \sum_{k=1}^{\infty} k \mathbf{D}_k$$

as $z = 1$. Also $e^{\mathbf{D}(1)t} \mathbf{e} = \mathbf{e}$. As a result, Equation 4.8 becomes:

$$\lambda = \pi \sum_{k=1}^{\infty} k \mathbf{D}_k \mathbf{e}. \tag{4.9}$$

Denote the arbitrary service time cumulative distribution function by $\bar{H}$ (.) with its finite mean μ^{-1}. Assume that the arrival process and the service process are independent and the traffic utilisation ρ satisfies:

$$\rho = \frac{\lambda}{\mu} < 1 \tag{4.10}$$

Define $\{\tau_n : n \geq 0\}$ as the successive epochs of departure with $\tau_0 = 0$. Let X_n and J_n denote the queue length and the phase of the arrival process at τ_n^+ Then $\{(X_n, J_n, \tau_{n+1} - \tau_n)\}$ form a semi-Markov sequence with state-space $\{0,1,...\} \times \{1,...m\}$ and a transition probability matrix $\bar{\mathbf{P}}(x)$. This sequence is referred to as a semi-Markov since the process may remain in a state with an arbitrary distribution of time and it is not constrained to the exponential distribution as is the case in ordinary Markov chains. At the instants of departure epochs the process, however, behaves like an ordinary Markov chain, hence at those instants an embedded Markov chain is formed. The transition probability matrix $\bar{\mathbf{P}}(x)$ of the semi-Markov process is given by:

$$\bar{\mathbf{P}}(x) = \begin{bmatrix} \bar{\mathbf{B}}_0(x) & \bar{\mathbf{B}}_1(x) & \bar{\mathbf{B}}_2(x) & . & . \\ \bar{\mathbf{A}}_0(x) & \bar{\mathbf{A}}_1(x) & \bar{\mathbf{A}}_2(x) & . & . \\ 0 & \bar{\mathbf{A}}_0(x) & \bar{\mathbf{A}}_1(x) & . & . \\ 0 & 0 & \bar{\mathbf{A}}_0(x) & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \tag{4.11}$$

where $\bar{\mathbf{A}}_n(x)$ and $\bar{\mathbf{B}}_n(x)$ are $m \times m$ matrices defined below and x denotes the time within which the next departure will take place. The matrices $\bar{\mathbf{A}}_n(x)$ and $\bar{\mathbf{B}}_n(x)$ are defined as:

$$\tilde{\mathbf{A}}_n(x) = \int_0^x \mathbf{P}(n,t)d\tilde{H}(t),\ n \geq 0,\ x \geq 0. \tag{4.12}$$

and

$$\tilde{\mathbf{B}}_n(x) = \sum_{v=1}^{n+1} \int_{0(u)}^{y} \int_{0(y)}^{x} e^{\mathbf{D}_0 u} \mathbf{D}_v P(n+1-v, y-u)d\tilde{H}(y-u)du \tag{4.13}$$

The matrix $\tilde{\mathbf{B}}_n(x)$ is obtained by conditioning on the time u and batch size v of the first arrival. Figure 4.4 clarifies the notation used.

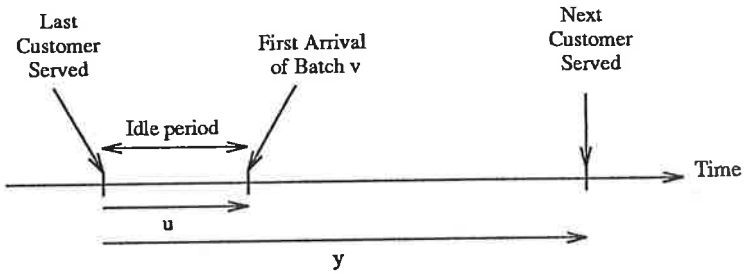


Figure 4.4: Notation for Evaluation of Matrix $\tilde{\mathbf{B}}_n(x)$

The matrices $\tilde{\mathbf{A}}_n(x)$ and $\tilde{\mathbf{B}}_n(x)$ can be interpreted as follows:

$[\tilde{A}_n(x)]_{jk}$ = P{given a departure at time 0, which left at least one customer in the queue and the arrival process in phase j, the next departure occurs no later than time x with the arrival process in phase k, and during that service there were n arrivals},

$[\tilde{B}_n(x)]_{jk}$ = P{given a departure at time 0, which left the queue empty and the arrival process in phase j, the next departure occurs no later than time x with the arrival process in phase k, and leaves n arrivals in the queue},

Define the following Laplace-Stieltjes transforms:

$$\mathbf{A}_n(s) = \int_0^\infty e^{-sx}\, d\tilde{\mathbf{A}}_n\,(x),$$

$$\mathbf{B}_n(s) = \int_0^\infty e^{-sx}\, d\tilde{\mathbf{B}}_n\,(x),$$

Also let

$$\mathbf{A}_n = \mathbf{A}_n(0) = \tilde{\mathbf{A}}_n(\infty),$$

$$\mathbf{B}_n = \mathbf{B}_n(0) = \tilde{\mathbf{B}}_n(\infty),$$

The stationary transition probability matrix of the Markov chain

$$\mathbf{P} = \tilde{\mathbf{P}}(\infty) = \begin{bmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \mathbf{B}_2 & \cdot & \cdot & \mathbf{B}_n & \cdot & \cdot \\ \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \cdot & \cdot & \mathbf{A}_n & \cdot & \cdot \\ 0 & \mathbf{A}_0 & \mathbf{A}_1 & \cdot & \cdot & \mathbf{A}_{n-1} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & \mathbf{A}_0 & \cdot & \cdot \end{bmatrix} \tag{4.14}$$

can be used to obtain the stationary queue-length distribution $\mathbf{x}$ at departure epochs. The defining system of equations is:

$$\mathbf{x}\,\tilde{\mathbf{P}}(\infty) = \mathbf{x}, \quad \mathbf{x}\,\mathbf{e} = 1. \tag{4.15}$$

Let $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \ldots)$ then we have:

$$\mathbf{x}_i = \mathbf{x}_0 B_i + \sum_{k=1}^{i+1} \mathbf{x}_k\, \mathbf{A}_{i-k+1}, \; i \geq 0. \tag{4.16}$$

by expanding $\tilde{\mathbf{P}}(\infty)$.

Note that x_{0j}, $1 \leq j \leq m$, is the stationary probability that a departure leaves the queue empty with the arrival process in phase j. In fact $(x_{0j})^{-1}$ is the mean recurrence time of the state $(0,j)$ in the Markov chain $\mathbf{P}$. Note also that given $\mathbf{x}_0$ the vectors $\mathbf{x}_i$ can be obtained recursively from Equation 4.16.

Next the stationary queue length distribution at an arbitrary time needs to be determined.

Define:

$$\mathbf{y}(i,j) = \lim_{t\to\infty} P\left\{X(t)=i, J(t)=j \mid X(0)=i', J(0)=j'\right\} \quad i \geq 0,\ 1 \leq j \leq m,\ t \geq 0. \tag{4.17}$$

$X(t)$ and $J(t)$ denote the queue length and the phase of the arrival process at time t.

Note that for $X(t)=i$ there must be some number of customers v in the queue just after the last departure epoch x and in the interval $(x,t]$ there must be $i-v$ arrivals.

Probability $K_{(i',j')(i,j)}(t) = P\left\{X(t)=i, J(t)=j \mid X(0)=i', J(0)=j'\right\}$ can be obtained as the solution of the Markov renewal equation [10] applied to the queue process as described in Appendix C.

The limiting behaviour of $K_{(i',j')(i,j)}(t)$ can be obtained by application of the key renewal theorem which is also described in Appendix C. As a result, the queue length distribution at an arbitrary time expressed in matrix notation becomes [46]:

$$\mathbf{y}_0 = -\lambda \mathbf{x}_0 \mathbf{D}_0^{-1} \tag{4.18a}$$

$$\mathbf{y}_i = \sum_{v=1}^{i} \lambda[\mathbf{x}_0 \mathbf{U}_v + \mathbf{x}_v] \int_0^\infty (1-\bar{H}(u)) P(i-v,u)\,du \quad \textit{for} \quad i \geq 1. \tag{4.18b}$$

where

$$\mathbf{U}_v = -\mathbf{D}_0^{-1}\,\mathbf{D}_v .$$

The (j,k)th entry of $\mathbf{U}_v$ denotes the conditional probability that an idle period ends with an arrival of group size v and the phase of the arrival process k, given that the idle period started with the arrival process in phase j.

4.3.2 Extension of Analysis for the BMAP/G/1/N Queue

The effect of a finite queue in the model is the truncation of the state space of the embedded Markov renewal process. If the queue is considered at the times of departure then the transition probability matrix $\tilde{\mathbf{P}}(x)$ of the semi-Markov chain formed by $\{(X_n, J_n, \tau_{n+1} - \tau_n)\}$ is given by:

$$\tilde{\mathbf{P}}(x) = \begin{bmatrix} \tilde{\mathbf{B}}_0(x) & \tilde{\mathbf{B}}_1(x) & \tilde{\mathbf{B}}_2(x) & \cdots & \tilde{\mathbf{B}}_{N-2}(x) & \sum_{k=N-1}^{\infty} \tilde{\mathbf{B}}_k(x) \\ \tilde{\mathbf{A}}_0(x) & \tilde{\mathbf{A}}_1(x) & \tilde{\mathbf{A}}_2(x) & \cdots & \tilde{\mathbf{A}}_{N-2}(x) & \sum_{k=N-1}^{\infty} \tilde{\mathbf{A}}_k(x) \\ 0 & \tilde{\mathbf{A}}_0(x) & \tilde{\mathbf{A}}_1(x) & \cdots & \tilde{\mathbf{A}}_{N-3}(x) & \sum_{k=N-2}^{\infty} \tilde{\mathbf{A}}_k(x) \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \tilde{\mathbf{A}}_0(x) & \sum_{k=1}^{\infty} \tilde{\mathbf{A}}_k(x) \end{bmatrix} \qquad (4.19)$$

where $\mathbf{A}_n(x)$ and $\mathbf{B}_n(x)$ are defined by Equations 4.12 and 4.13. The stationary transition probability matrix of the Markov chain $P = \tilde{\mathbf{P}}(\infty)$ is given by:

$$\mathbf{P} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{B}_1 & \mathbf{B}_2 & \cdots & \mathbf{B}_{N-2} & \sum_{k=N-1}^{\infty} \mathbf{B}_k \\ \mathbf{A}_0 & \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{N-2} & \sum_{k=N-1}^{\infty} \mathbf{A}_k \\ 0 & \mathbf{A}_0 & \mathbf{A}_1 & \cdots & \mathbf{A}_{N-3} & \sum_{k=N-2}^{\infty} \mathbf{A}_k \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{A}_0 & \sum_{k=1}^{\infty} \mathbf{A}_k \end{bmatrix} \qquad (4.20)$$

Explicit expressions of the probabilities $P_{ij}(n,t)$ are difficult to obtain except for simple cases like Poisson, however an efficient algorithm for the computation of

matrices $\mathbf{A}_n$ has been given in [37]. The scheme is outlined in Appendix D.

The matrices $\mathbf{B}_n$, defined in Equation 4.13 in Section 4.3.1, are given in the stationary case by:

$$\mathbf{B}_n = -\mathbf{D}_0^{-1} \sum_{k=0}^{n} \mathbf{D}_{k+1}\mathbf{A}_{n-k}. \tag{4.21}$$

This expression can be obtained from Equation 4.13 by evaluating the Laplace-Stieltjes transform $\mathbf{B}_n(s)$ at $s = 0$ as presented in [38].

The computation of the stationary probability vector $\mathbf{x}$ at departure epochs can be carried out as described in Appendix E.

After computation of matrix $\mathbf{x}$, the queue length distribution at an arbitrary time instant t is required. Let $\mathbf{y} = (\mathbf{y}_0, \ldots, \mathbf{y}_N)$. The vector $\mathbf{y}_0$ can be obtained by carrying out analysis analogous to that presented for the infinite queue, where the key renewal theorem was applied. In this case, however, the renewal points corresponding only to the departure epochs from an empty queue have to be considered. Then [6]:

$$\mathbf{y}_0 = [\mu^{-1} - \mathbf{x}_0\mathbf{D}_0^{-1}\mathbf{e}]^{-1}[-\mathbf{x}_0\mathbf{D}_0^{-1}] \tag{4.22}$$

As an example, the above formula reduces to:

$y_0 = (1 - \rho)$ for $M/G/1$

and

$y_0 = x_0(\rho + x_0)^{-1}$ for $M/G/1/N$

where $\rho = \lambda/\mu$.

The probability that the server is idle is given by:

$$\begin{aligned} P_{idle} &= \mathbf{y}_0\,\mathbf{e} \\ &= -\mathbf{x}_0\,\mathbf{D}_0^{-1}\mathbf{e}[\mu^{-1} - \mathbf{x}_0\mathbf{D}_0^{-1}\mathbf{e}]^{-1} \end{aligned} \tag{4.23}$$

and the probability that the server is busy is given as:

$$\begin{aligned} P_{busy} &= 1 - P_{idle} \\ &= \mu^{-1}[\mu^{-1} - \mathbf{x}_0\mathbf{D}_0^{-1}\mathbf{e}]^{-1} \end{aligned} \tag{4.24}$$

The queue length distribution $\mathbf{y}_i$ for $1 \le i \le N$ can be found using the supplementary variable technique [11], [36], [30]. The use of this technique for this case is described in Appendix G. This leads to the following expression for $\mathbf{y}_i$:

$$\mathbf{y}_i = \frac{1}{\mu^{-1} - \mathbf{x}_0\mathbf{D}_0^{-1}\mathbf{e}} \left[\mathbf{x}_0[\mathbf{D}_{i-1}(0) + \sum_{\nu=1}^{i} (\mathbf{D}_0^{-1}\mathbf{D}_\nu)\mathbf{D}_{i-\nu}(0)] + \sum_{\nu=1}^{i-1} \mathbf{x}_0[\mathbf{D}_{i-\nu-1}(0) - \mathbf{D}_{i-\nu}(0)] - \mathbf{x}_i\,\mathbf{D}_0(0) \right] \tag{4.25}$$

for $1 \le i \le N-1$. Also $\mathbf{D}_n(s)$ can be obtained as the coefficients of z^n from the expansion of:

$$[\mathbf{D}(z) + s\mathbf{I}]^{-1} = \sum_{n=0}^{\infty} \mathbf{D}_n(s) z^n$$

Now, the blocking probability can be obtained from:

$$P_{block} = 1 - \sum_{n=0}^{N-1} \mathbf{y}_n\,\mathbf{e} \tag{4.26}$$

4.4 Algorithm for the Solution of the two-state MMPP/D/1/N Queue

In this section the expressions presented in the previous sections will be used to produce an algorithm for obtaining the queue length distribution at an arbitrary time and hence the blocking probability for the two-state MMPP/D/1/N queue.

The main steps of the algorithm are listed below. Each step, which is based on the proceeding theory, is discussed in detail.

The Algorithm

The main steps of the algorithm for the solution of the two-state MMPP/D/1/N queue are shown in Figure 4.5.

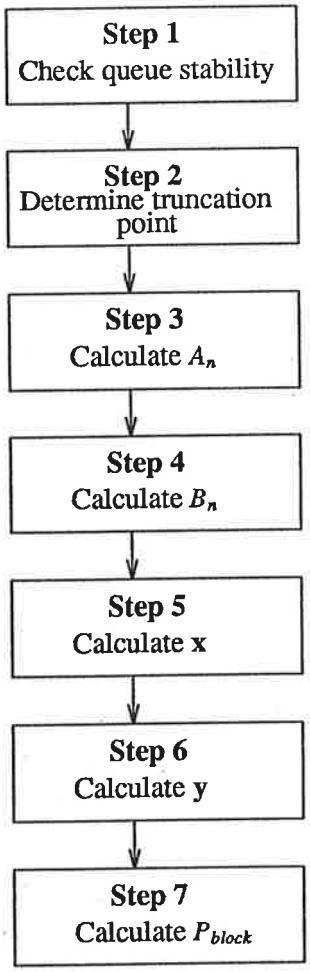


Figure 4.5: 2-state MMPP/D/1/N Queue Algorithm Steps

Assume that the parameters describing the two-state MMPP (i.e. $\lambda_1, \lambda_2, r_1$, and r_2) are given. Also given is the service rate d and the depth of the queue N.

Step 1: Check the stability condition.

For the queue to be stable we require that:

$$\rho = \frac{\lambda}{\mu} < 1$$

where μ^{-1} is the mean service time and λ^{-1} is the mean interarrival time. For a two-state MMPP:

$$\lambda^{-1} = \frac{r_1+r_2}{r_2\lambda_1+r_1\lambda_2}$$

and for deterministic service distribution $\mu^{-1} = d^{-1}$ where d denotes the deterministic service rate. Hence the stability condition for the two-state MMPP/D/1/N queue is:

$$\rho = \frac{\lambda_1 r_2+\lambda_2 r_1}{r_1+r_2}\, d^{-1} < 1 \qquad (4.27)$$

Step 2: Determine the truncation point for the computation of transition probability matrices $\mathbf{A}_n$.

As detailed in Step 3, matrices $\mathbf{A}_n$ can be computed from :

$$\mathbf{A}_n = \sum_{l=n}^{\infty} \gamma_l \mathbf{K}_n^l$$

where the sequence γ_n is described in Appendix D.

Since $\sum_{n=0}^{\infty} \gamma_n = 1$ the choice of the truncation index N_0 is given by the smallest index N_0 for which

$$\sum_{n=0}^{N_0} \gamma_n = 1 - \varepsilon$$

The choice of $\varepsilon = 10^{-8}$ is recommended in [47] and [23].

For the two-state MMPP, γ_l is given by [25]:

$$\gamma_l = \int_0^\infty e^{-\theta t} \frac{(\theta t)^l}{l!} d\bar{H}(t)$$

where

$$\theta = \max(\lambda_j - R_{jj})$$

where

$$\mathbf{R} = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}$$

So we have:

$$\theta = \max(\lambda_1 + r_1, \lambda_2 + r_2)$$

When the service distribution $\bar{H}(.)$ is deterministic with mass at d^{-1} then γ_n can be computed recursively from [25]:

$$\gamma_0 = e^{-\theta d^{-1}} \tag{4.28a}$$

$$\gamma_n = (\frac{\theta d^{-1}}{n})\gamma_{n-1}. \tag{4.28b}$$

Step 3: Calculate transition probability matrices $\mathbf{A}_n$.

In this step the computation of the sequence $\{\mathbf{A}_0,..,\mathbf{A}_{N_0}\}$ is carried out with N_0 as determined in Step 2. The sequence $\{\gamma_0,..,\gamma_{N_0}\}$ will also be used.

The elements $\mathbf{A}_n$ can be computed from:

$$\mathbf{A}_n = \sum_{l=n}^{\infty} \gamma_l \mathbf{K}_n^l \tag{4.29}$$

as described in Appendix D.

For a two-state MMPP the elements $\mathbf{K}_n^l$ are given as follows:

$$\mathbf{K}_0^0 = \mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{K}_0^{l+1} = \mathbf{K}_0^l (\mathbf{I} + \theta^{-1} (\mathbf{R} - \Lambda))$$

where **R** is as given in Step 2 and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

Also

$$\mathbf{K}_n^0 = 0 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$\mathbf{K}_n^{l+1} = \mathbf{K}_n^l (\mathbf{I} + \theta\Lambda) + \mathbf{K}_{n-1}^l \theta \Lambda$$

Taking into account the truncation index N_0, the elements of $\mathbf{K}_n$ can be calculated from:

$$\mathbf{A}_n = \sum_{l=n}^{N_0} \gamma_l \mathbf{K}_n^l + \left(\sum_{l=N_0+1}^{\infty} \gamma_l \right) \mathbf{K}_n^{N_0} \tag{4.30}$$

Step 4: Calculate transition probability matrices $\mathbf{B}_n$.

The elements $\mathbf{B}_n$ are calculated (see Equation 4.21) by:

$$\mathbf{B}_n = \sum_{v=1}^{n+1} (-\mathbf{D}_0^{-1} \mathbf{D}_v) \mathbf{A}_{n-v+1} \tag{4.31}$$

where

$$(-\mathbf{D}_0^{-1} \mathbf{D}_1) = \left\{ \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} - \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix} \right\}^{-1} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

and $\mathbf{D}_v = 0$ for $v > 1$

Step 5: Calculate the queue length distribution at departure epochs **x**.

The inverse matrix $\mathbf{T}^{-1}$ is necessary for the calculation of the vector **x**. As described in Appendix E:

$$\mathbf{T} = \begin{bmatrix} \mathbf{E}_0 & \mathbf{E}_1 & \mathbf{E}_2 & . & \mathbf{E}_{N-2} \\ 0 & \mathbf{E}_0 & \mathbf{E}_1 & . & \mathbf{E}_{N-3} \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & \mathbf{E}_0 \end{bmatrix} = \begin{bmatrix} -\mathbf{A}_0 & \mathbf{I}-\mathbf{A}_1 & -\mathbf{A}_2 & . & -\mathbf{A}_{N-2} \\ 0 & -\mathbf{A}_0 & \mathbf{I}-\mathbf{A}_1 & . & -\mathbf{A}_{N-3} \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & 0 & -\mathbf{A}_0 \end{bmatrix} \tag{4.32}$$

The inverse of this matrix can be obtained by following the algorithm for inverses of Toeplitz matrices as presented in Appendix F.

Having calculated $\mathbf{T}^{-1}$ its Schur complement Δ can be obtained from (see Appendix E):

$$\Delta = \mathbf{B} - \mathbf{C}\mathbf{T}^{-1}\mathbf{D}$$

where

$$\mathbf{B} = \mathbf{F}_{N-1} = \begin{bmatrix} -\sum_{k=N-1}^{N_q} (B_{00})_k & 1 \\ -\sum_{k=N-1}^{N_q} (B_{10})_k & 1 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} -\sum_{k=N-1}^{N_q} (A_{00})_k & -\sum_{k=N-1}^{N_q} (A_{10})_k & -\sum_{k=N-2}^{N_q} (A_{00})_k & -\sum_{k=N-2}^{N_q} (A_{10})_k & . . & -\sum_{k=1}^{N_q} (A_{00})_k & -\sum_{k=1}^{N_q} (A_{10})_k \\ 1 & 1 & 1 & 1 & . . & 1 & 1 \end{bmatrix}^t$$

and

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}-\mathbf{B}_0 & -\mathbf{B}_1 & -\mathbf{B}_2 & -\mathbf{B}_3 & . . & -\mathbf{B}_{N-3} & -\mathbf{B}_{N-2} \end{bmatrix}$$

The vector $\mathbf{x}$ can be obtained as the last row of the matrix:

$$\begin{bmatrix} \Delta^{-1} & -\Delta^{-1}\mathbf{F} \end{bmatrix}$$

where $\mathbf{F} = \mathbf{CT}$.

Step 6: Calculate the queue length distribution at arbitrary times **y**.

In this step vector $\mathbf{y}_n$ describing the queue length distribution is obtained. For the two-state MMPP/D/1/N queue the expressions given for $\mathbf{y}_0$ in Equation 4.22 and $\mathbf{y}_n$ in Equation 4.25 of Section 4.3.2 become:

$$\mathbf{y}_0 = \frac{-\mathbf{x}_0\mathbf{D}_0^{-1}}{[\mu^{-1} - \mathbf{x}_0\mathbf{D}_0^{-1}]} \tag{4.33a}$$

$$\mathbf{y}_n = \frac{1}{[\mu^{-1} - \mathbf{x}_0\mathbf{D}_0^{-1}]}\Big[\mathbf{x}_0[\mathbf{D}_{n-1}(0) + (\mathbf{D}_0^{-1}\mathbf{D}_1)\mathbf{D}_{n-1}(0)] + \sum_{\nu=1}^{n-1}\mathbf{x}_\nu[\mathbf{D}_{n-\nu-1}(0) - \mathbf{D}_{n-\nu}(0)] - \mathbf{x}_n\mathbf{D}_0(0)\Big] \tag{4.33b}$$

where

$$\mathbf{D}_0 = \begin{bmatrix} -r_1 - \lambda_1 & r_1 \\ r_2 & -r_2 - \lambda_2 \end{bmatrix}$$

$$\mathbf{D}_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

and

$$\mathbf{D}_n(0) = -\mathbf{D}_0^{-1}(\mathbf{D}_1\mathbf{D}_0^{-1})^n$$
$$= -(\mathbf{R} - \Lambda)^{-1}(\Lambda(\mathbf{R} - \Lambda)^{-1})^n.$$

Step 7: Calculate the blocking probability P_{block}.

The blocking probability for the two-state MMPP/D//1/N queue can be finally obtained from:

$$P_{block} = 1 - \sum_{n=0}^{N-1}\mathbf{y}_n\mathbf{e}. \tag{4.34}$$

The algorithm for the solution of the two-state MMPP/D/1/N queue presented above was implemented in software using the C++ programming language. The

algorithm required 2000 lines of code.

4.5 Conclusion

In this chapter a queueing model for the connection admission control in ATM networks was analysed. The queueing system was represented as a two-state MMPP/D/1/N queue. After presenting a more general BMAP process, of which a two-state MMPP is a special case, BMAP/G/1/∞ queue was analysed, followed by its finite equivalent: the BMAP/G/1/N queue. This led to an algorithmic solution of the two-state MMPP/D/1/N queue from which its loss probability was obtained. This loss probability will be used in subsequent chapters as a quality of service measure for connections in ATM networks and hence as the admission criterion in the connection admission algorithm which is presented in Chapter 6.

Chapter Five:

Traffic Superposition in Stream Arrival Model

5.1 Introduction

The model of connection admission control proposed in Chapter 3 was separated into two parts. The first part, involving the queueing model for the virtual path, was analysed in Chapter 4. In the current chapter the problem of obtaining the aggregate stream is addressed. Individual connections which share the same virtual path must be combined into a single stream representing total traffic carried in the virtual path.

In Section 5.2 general approaches to traffic superposition, which have been proposed in the literature, are discussed. Two particular methods for stream superposition developed in [24] and [51] respectively are studied in more detail. The two methods are compared in Section 5.3 and as a result one method is proposed for use in the connection admission algorithm. Section 5.4 presents the superposition algorithm and in Section 5.5 superposition of heterogeneous traffic streams is studied.

5.2 Stream Superposition

The problem of stream superposition is closely related to the problem of traffic characterisation. It is not clear at this time [53], [54] which characteristics in a stream should be chosen for use in a canonical process which will then accurately

predict queueing performance. It may also be difficult to measure the required statistical characteristics.

A common approach to superposition of traffic streams in the past was to approximate the combined stream by a Poisson process. This approach assumed convergence to a Poisson process when the number of streams to be superimposed was large. For the convergence to behave in this manner, however, it was required that as the number of streams increased each stream should contribute less and less so that the total intensity was kept constant. This 'thining' condition was often difficult to satisfy resulting in inaccurate results [26].

Matching of moments of the superposition process to the moments of a canonical process has been used extensively in the superposition approximation. For example, the equivalent random method, which has been used for superposition of traffic streams offered to a blocking system, matches the mean and peakedness of the superposition to the mean and peakedness of a canonical process. The concept of peakedness has been defined in Section 2.3.2 of Chapter 2 (see Equation 2.1).

Another form of characterisation of a superposition process has been based on the interarrival time coefficient of variation. Two common approaches here have been to consider the stationary and asymptotic intervals [60]. In the stationary interval method the distribution of the interval in the approximating renewal process is the same as the distribution of an interval in the superposition process, whereas in the asymptotic interval method the moments of the approximating renewal process over a large time interval are matched with the moments of the superposition process over a large time interval. These two methods have been also combined to produce a hybrid method to capture the effect of the superposition process over a wide range of operating conditions of a queue. The hybrid coefficient of variation is a combination of the asymptotic and stationary interval coefficients of variation and is given by [1]:

$$c_h^2 = wc_a^2 + (1-w)c_s^2$$

where w is a function of traffic intensity and the number of component processes. Different canonical renewal processes have been used depending on the value of c_h^2.

When the complexity of the superposition stream increases, as the bursts present in component streams do not disappear in the superposition process, renewal approximations may become inadequate [26]. To overcome this problem the superposition process has been approximated by a correlated non-renewal stream chosen in such a way so that several of its statistical characteristics match those of the original superposition. The canonical process in this case is the Markov modulated Poisson process. In particular, a two-state version of this process has been used extensively for this purpose. The main characteristics of the two-state MMPP have been presented in Section 4.2 of Chapter 4. The process is characterised by four parameters which can be chosen to match different characteristics of the superposition process.

Amongst the methods which have been proposed for matching characteristics of a real life stream to the parameters of the two-state MMPP canonical stream include [63]:

1. Matching the first three moments and the integral of the covariance function of the arrival rate [24],

2. Matching the first two moments of the interarrival times, the asymptotic variance to mean ratio and the asymptotic covariance of the number of arrivals [51],

3. Matching the first three moments of the number of arrivals in a finite time interval and the asymptotic variance to mean ratio of the number of arrivals [25] (In this method correlation properties of the superposition stream are approximated by choosing a process which provides a good match to the variance-time curve),

4. Estimating the parameters of the two-state MMPP using an iterative procedure which is based on the maximum likelihood estimation [41].

In the following subsections the first two methods will be used for obtaining superpositions of two-state MMPPs. They will be presented in more detail and compared in order to choose one of them for use in a connection admission algorithm.

5.2.1 Method I for Superposition of 2-state MMPPs

Method I which was developed in [24] uses the following set of characterising parameters to describe a traffic stream:

m - mean arrival rate,

v - variance of arrival rate,

μ - third moment of arrival rate,

τ - time constant for the arrival process as defined below.

The above parameters are related to the defining parameters of the two-state MMPP in the following manner:

$$m = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} \tag{5.1}$$

$$v = \frac{r_1 r_2 (\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2} \tag{5.2}$$

$$\mu = \frac{\lambda_1^3 r_2 + \lambda_2^3 r_1}{r_1 + r_2} \tag{5.3}$$

$$\tau = \frac{1}{v}\int_0^\infty r(t)dt = \frac{1}{r_1 + r_2} \tag{5.4}$$

where r(t) is the covariance function of the arrival rate and it is given for the two-state MMPP by:

$$r(t) = \frac{r_1 r_2 (\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2} e^{-(r_1 + r_2)t}$$

When n two-state MMPP streams are superimposed together characterising parameters are then given by:

$$m = \sum_{i=1}^{n} m_i \tag{5.5}$$

$$\nu = \sum_{i=1}^{n} \nu_i \tag{5.6}$$

$$\mu^* = \sum_{i=1}^{n} \mu_i^* \tag{5.7}$$

$$\tau = \sum_{i=1}^{n} \frac{\nu_i}{\nu} \tau_i \tag{5.8}$$

where

$$\mu^* = \mu - 3m\nu - m^3$$

Note that the above equations for obtaining the characterising parameters of the superimposed stream from the characterising parameters of the component streams do not depend on the assumption that the component streams are themselves modelled by two-state MMPPs.

The defining parameters for the two-state MMPP representing the superimposed stream can be obtained by inversion of the above equations resulting in [24]:

$$r_1 = \frac{1}{\tau(1+\eta)} \tag{5.9}$$

$$r_2 = \frac{\eta}{\tau(1+\eta)} \tag{5.10}$$

$$\lambda_1 = m + \sqrt{\nu/\eta} \tag{5.11}$$

$$\lambda_2 = m - \sqrt{\nu\eta} \tag{5.12}$$

where

$$\eta = 1 + \frac{\delta}{2}[\delta - \sqrt{4+\delta^2}]$$

and

$$\delta = \frac{\mu_3^*}{\nu^{3/2}}$$

5.2.2 Method II for Superposition of Two-state MMPPs

In Method II which was developed by Rossiter [51], the set of characterising parameters which characterises the arrival process consists of:

λ - mean arrival rate,

c^2 - squared coefficient of variation of interarrival times,

$Z(\infty)$ - asymptotic variance to mean ratio of the number of arrivals,

$C(\infty)$ - asymptotic covariance of the number of arrivals.

The squared coefficient of variation c^2 of a process is given by:

$$c^2 = \frac{var(X)}{[E(X)]^2}$$

where X denotes the interarrival time between consecutive arrivals. As a result c^2 is given by variance divided by the square of the mean of the interarrival time. Note that for a Poisson process $var(X) = \lambda^{-2}$ and $E(X) = \lambda^{-1}$ giving $c^2 = 1$.

Asymptotic variance to mean ratio of the number of arrivals $Z(\infty)$ is given by:

$$Z(\infty) = \lim_{t \to \infty} Z(t)$$

where

$$Z(t) = \frac{var[N(t)]}{E[N(t)]}$$

where N(t) is the number of arrivals during time t. For a Poisson process $var[N(t)] = \lambda t$ and $E[N(t)] = \lambda t$ giving $Z(t) = 1$ for all t.

The asymptotic covariance of the number of arrivals $C(\infty)$ is given by:

$$C(\infty) = \lim_{t \to \infty} C(t)$$

where C(t) is the covariance of the number of arrivals in adjacent intervals of length t and it is given by [51]:

$$C(t) = cov[N(t), N(2t) - N(t)] = \frac{1}{2} var[N(2t)] - var[N(t)]$$

The characterising parameters can be obtained from the defining parameters of the two-state MMPP as [51]:

$$\lambda = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} \tag{5.13}$$

$$c^2 = 1 + \frac{2 r_2 r_1 (\lambda_1 - \lambda_2)^2}{(\lambda_1 r_2 + \lambda_2 r_1 + \lambda_1 \lambda_2)(r_1 + r_2)^2} \tag{5.14}$$

$$Z(\infty) = 1 + \frac{2 r_2 r_1 (\lambda_1 - \lambda_2)^2}{(\lambda_1 r_2 + \lambda_2 r_1)(r_1 + r_2)^2} \tag{5.15}$$

$$C(\infty) = \frac{r_1 r_2 (\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^4} \tag{5.16}$$

When n two-state MMPP streams are superimposed together the set of characterising parameters for the superimposed set is given by [51]:

$$\lambda = \lambda_1 + \cdots + \lambda_n \tag{5.17}$$

$$Z(\infty) = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} Z_i(\infty) \tag{5.18}$$

$$C(\infty) = \sum_{i=1}^{n} C_i(\infty) \tag{5.19}$$

$$c^2 = E(T) 2\lambda - 1 \tag{5.20}$$

where E(T) is the mean of the forward recurrence time which is given by:

$$E(T) = \int_0^\infty P_N(0,t)\,dt \tag{5.21}$$

where $P_N(z,t)$ is the probability generating function of the number of arrivals during an interval (0,t]. For superposition of n streams we have that:

$$P_N(z,t) = \prod_{i=1}^{n} P_{N_i}(z,t) \tag{5.22}$$

which assumes that individual streams are independent and stationary.

For an individual two-state MMPP stream $P_N(z,t)$ is given by [51]:

$$P_N(z,t) = \frac{1}{\sqrt{\Delta}} \left[[r_1 + r_2 + \frac{\lambda_1 r_1 + \lambda_2 r_2}{r_1 + r_2}(1 - z) - \eta_1] e^{-\eta_1 t} - [r_1 + r_2 + \frac{\lambda_1 r_1 + \lambda_2 r_2}{r_1 + r_2}(1 - z) - \eta_2] e^{-\eta_2 t} \right] \tag{5.23}$$

where

$$\eta_1 = \frac{1}{2}[(\lambda_1 + \lambda_2) + r_1 + r_2 - \sqrt{\Delta}]$$

$$\eta_2 = \frac{1}{2}[(\lambda_1 + \lambda_2) + r_1 + r_2 + \sqrt{\Delta}]$$

$$\Delta = [(\lambda_1 - \lambda_2) - r_2 + r_1]^2 + 4r_1r_2$$

From the above expressions it can be seen that the calculation of the coefficient of variation c^2 is rather involved. In order to obtain $E(T)$ needed in the calculation, the product of probability generating functions $P_{N_i}(z,t)$ of each individual stream must be obtained. The resulting expression $P_N(z,t)$ evaluated at $z = 0$, which is a linear combination of negative exponential terms, is then integrated to produce $E(T)$. The above calculations must also be carried out using all individual streams when an additional stream is added to the superposition. This fact has important implications on the performance of any connection admission algorithm which is to use this superposition method. The complexity of these calculations can, however, be reduced when all individual streams are identical.

There is a one-to-one correspondence between the sets of defining and characterising parameters. The set of defining parameters $(\lambda_1, \lambda_2, r_1, r_2)$ for the two-state MMPP representing the superimposed stream can be obtained from the set of characterising parameters $(\lambda, c^2, Z(\infty), C(\infty))$ [51] after some straightforward but tedious algebraic manipulation. More details of the algebra involved are included in Appendix H together with expressions for the defining parameters.

5.3 Superposition Methods Testing and Comparison Results

The following steps are necessary in order to obtain the superposition of n two-state MMPP streams using the two methods described above:

Method I

i. From the defining parameters of each component stream obtain the corresponding characterising parameters according to Equations 5.1-5.4.

ii. Obtain the characterising parameters of the superimposed stream according to Equations 5.5-5.8.

iii. Obtain the defining parameters of the two-state MMPP approximating the superposition stream according to Equations 5.9-5.12.

Method II

i. Obtain the characterising parameters of component streams from the corresponding defining parameters according to Equations 5.13-5.16.

ii. Obtain the characterising parameters of the superimposed stream according to Equations 5.17-5.20. This step also involves calculation of the probability generating function for each component stream as given in Equation 5.23, followed by the probability generating function for the superposition (Equation 5.22) and the integration in Equation 5.21.

iii. Obtain the defining parameters of the two-state MMPP approximating the superimposed stream as shown in Appendix H.

The two superposition methods were implemented in software using the C++ programming language. Method I required 300 lines of code and Method II required 800 lines of code.

In order to determine the suitability of the two superposition methods for connection admission control, simple tests were carried out as described below.

An arrival model based on the voice model presented in [56] was chosen with the following parameters: $\lambda_1 = 54.44$ cells/sec, $\lambda_2 = 0$, $r_1 = 2.268\ \text{sec}^{-1}$, and $r_2 = 1.532\ \text{sec}^{-1}$. Thus the voice process was represented by an interrupted Poisson

process (IPP), which is a special case of the two-state MMPP. The model is also depicted in Figure 5.1.

Using the two methods, n such voice streams were superimposed together. The characterising parameters of Method II (i.e. the coefficient of variation c^2, mean arrival rate m, asymptotic variance to mean ratio $Z(\infty)$ and asymptotic covariance $C(\infty)$ of the number of arrivals in the superimposed stream) were used for comparison between the two methods. For Method I these characterising parameters were obtained from the corresponding defining parameters of the two-state MMPPP approximating the superimposed stream. Figures 5.2 and 5.3 show the results of these tests. The values of the defining parameters of the two-state MMPP for the superimposed stream obtained using both methods are also compared as shown in Figures 5.4-5.7.

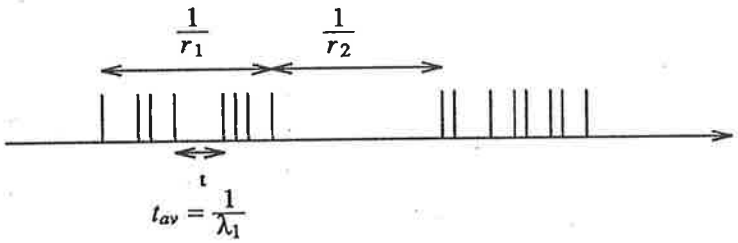


Figure 5.1: An IPP Model for a Voice Stream

The value of the coefficient of variation c^2 obtained using Method II is exact as it is used in the matching process. Thus the curve for c^2 obtained using Method I can be regarded as a test of how closely this method approximates the coefficient of variation for the superimposed stream. As can be seen in Figure 5.2 the coefficient of variation obtained using Method I approaches that of Method II as the number of streams reaches $n \approx 10$. The asymptotic variance to mean ratio of the number of arrivals for the superposition, which is also shown in Figure 5.2, can be seen to match using both methods.

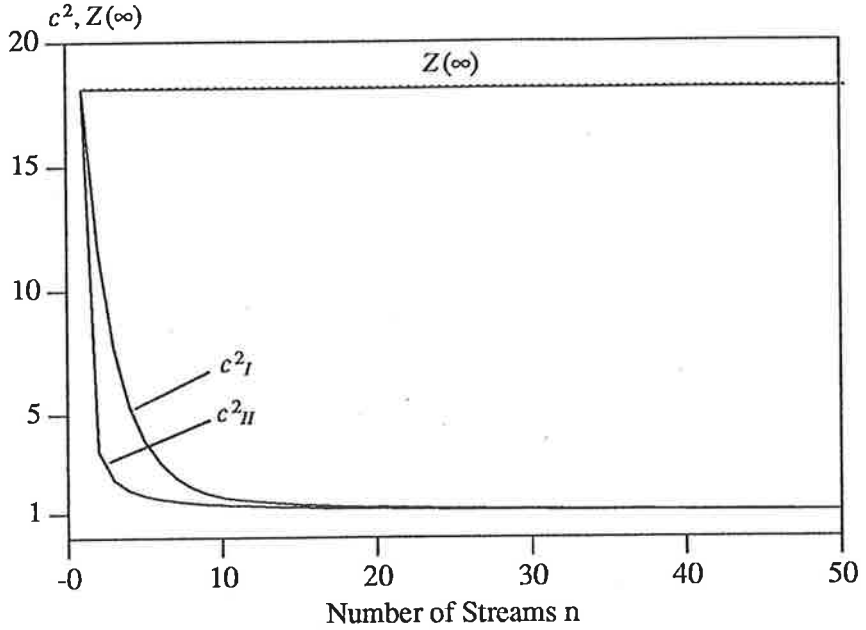


Figure 5.2: Coefficient of Variation c^2 and Asymptotic Variance to Mean Ratio $Z(\infty)$ for Stream Superposition

The mean arrival rate m and the asymptotic covariance of the number of arrivals $C(\infty)$ in the superimposed stream also match for the two methods as can be seen in Figure 5.3. This is expected in the case of the mean as both methods use it to characterise the process. The matching of the covariance values, however, can be due to the time constant of the arrival process τ used in Method I accurately reflecting the asymptotic covariance nature of the process.

Note also that in order for the superposition to approach a Poisson process it is required that the coefficient of variation and the asymptotic variance to mean ratio approach the value of 1, and the asymptotic covariance approaches the value of 0. That is, it is required that $c^2 = Z(\infty) = 1$ and $C(\infty) = 0$. From Figures 5.2 and 5.3 it can been seen that the superposition becomes further from a Poisson process as the number of streams n increases. This is because individual processes do not contribute less and less to the superposition as required from the classical theorem

[26]. For this to happen the intensity of the superimposed stream would have to be kept constant. Note that $Z(\infty)$ remains constant and independent of n. This is also in agreement with Proposition 2 in [56].

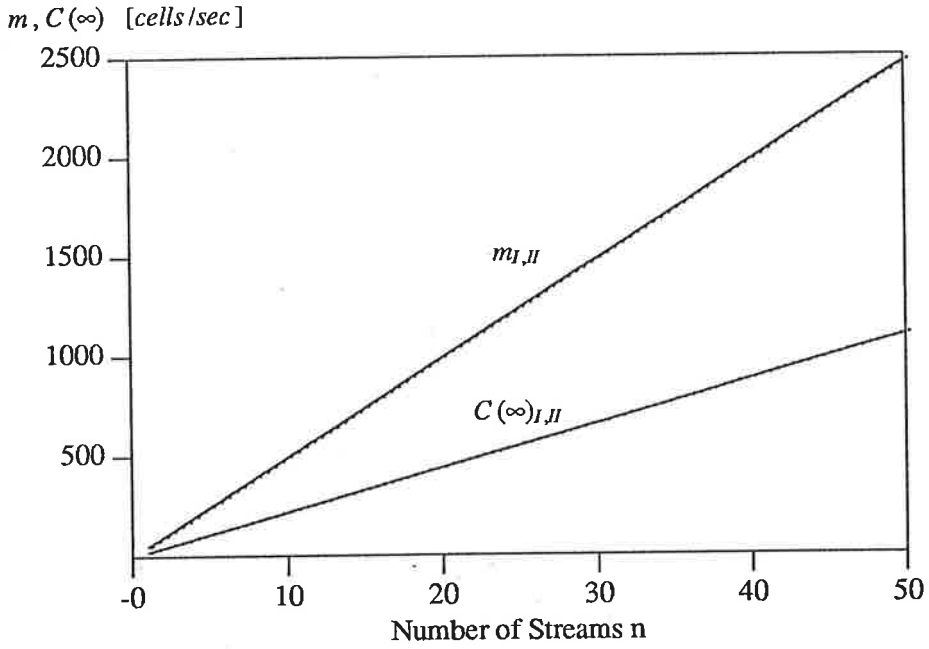


Figure 5.3: Mean m and Covariance $C(\infty)$ for Stream Superposition

In order for the superposition method to be useful for the connection admission algorithm it should be easy to implement. One of the important considerations in the implementation of the two methods was the possibility of carrying out the process in stages. This would avoid having to completely decompose an existing superimposed stream into individual streams in order to superimpose an additional stream to the existing aggregate. In order to study the effect of stream grouping on the superposition when using the two methods, arrival streams modelled as above were superimposed individually up to 10 streams. Subsequently, when additional streams were added the superposition of 10 streams was regarded as a single stream. For example, superposition of 36 streams was carried out as: 3 streams representing superposition of 10 individual streams plus 6 individual streams. The effect of such stream grouping was studied by observing its effect on the phase ratio θ $(= \frac{r_1}{r_2})$,

components rates λ_i, and component phase rates r_i of the two-state MMPP representing the superimposed stream. Method I gives the same answer regardless of stream grouping used.

The effect of stream grouping on the phase ratio θ using Method II can be seen in Figure 5.4. In addition it can be seen that the phase ratios obtained using the two methods vary considerably. By studying the absolute values of component rates λ_i and r_i further insight can be gained.

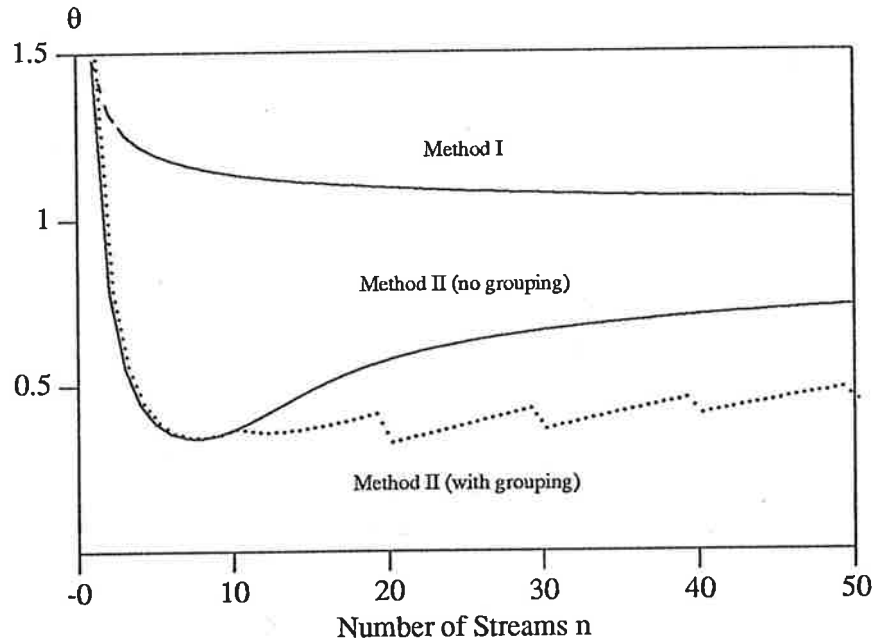


Figure 5.4: Effect of Stream Grouping on Phase Ratio θ

Component rates λ_i ($i = 1,2$) obtained using the two methods are shown in Figure 5.5. It can be seen that the two rates are always greater when using Method I as compared to those obtained from Method II. However, as was noted in Figure 5.3 the mean arrival rates of the two methods matched. As a result, the phase rates r_i ($i = 1,2$) obtained using the two methods should compensate for the difference in the component rates. This is in fact the case as can be seen from Figure 5.6. In this figure the phase rates r_i are compared for the two methods. Phase rates for Method II were obtained in this case without stream grouping. It can also be seen that the

values of the phase rate tend to $r = \frac{r_1 + r_2}{2}$ as the number of streams n increases. This happens much faster for Method I than for Method II.

The effect of stream grouping on the phase rates when using Method II is shown in Figure 5.7. It can be seen from the figure that stream grouping reduces the rate of convergence of the two phase rates r_1 and r_2.

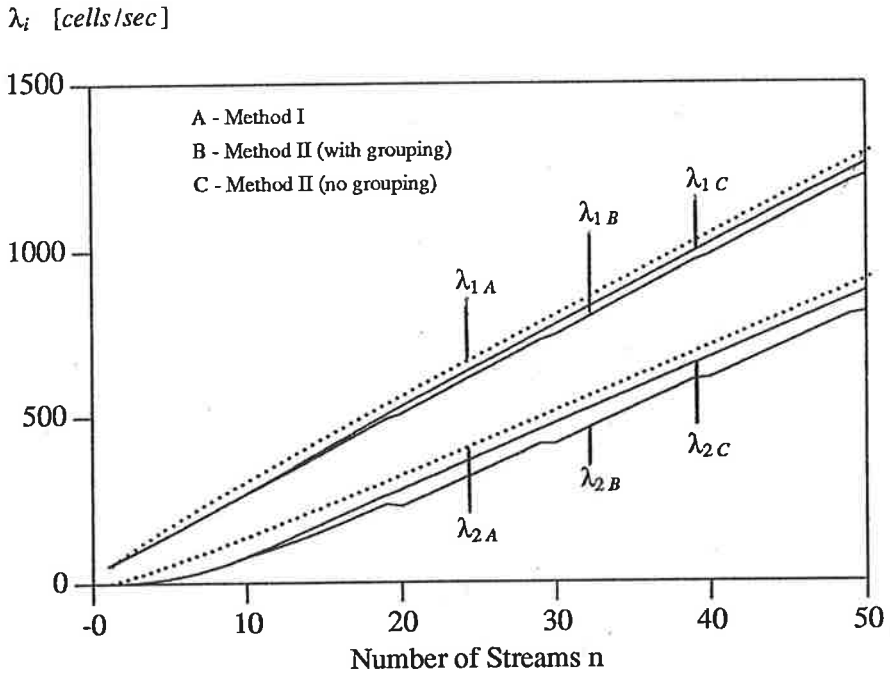


Figure 5.5: Effect of Stream Grouping on Component Rates λ_i

From the figures presented it appears that the parameters of the two-state MMPP approximating the superposition of a number of two-state MMPPs are quite sensitive to stream grouping when Method II is used. This is particularly evident in the case of phase rates r_i as shown in Figure 5.7. This sensitivity, however, does not give any indication of how well the superimposed stream will predict the queueing performance of the real life traffic. However, one important consequence of avoiding stream grouping when using Method II would be the elimination of the expensive computational cost associated with this superposition method. The expensive step in this method is associated with the calculation of the coefficient of variation c^2. In particular, the computational cost increases as 2^n as the number of

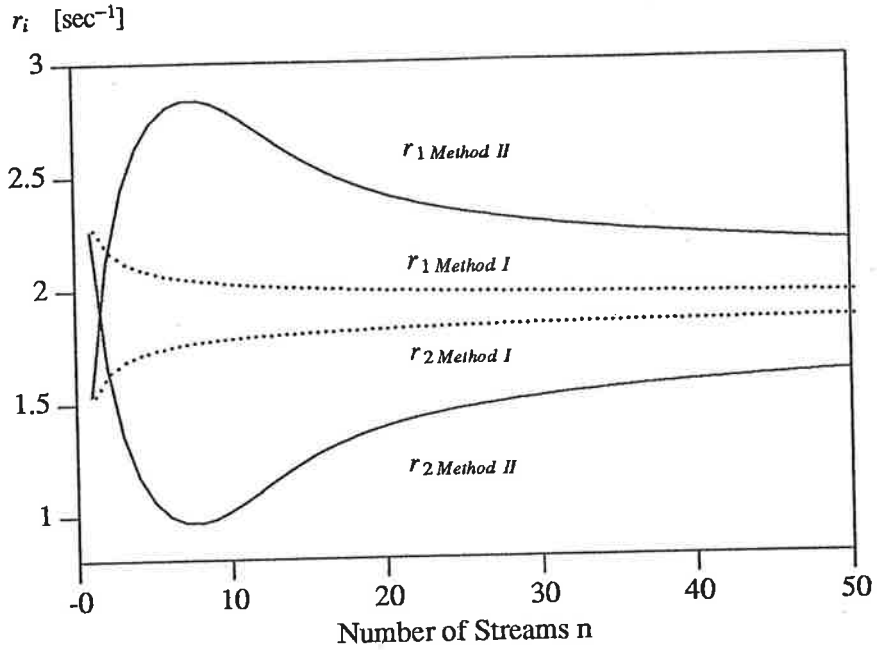


Figure 5.6: Phase Rates r_i for Stream Superposition Using Method I and II

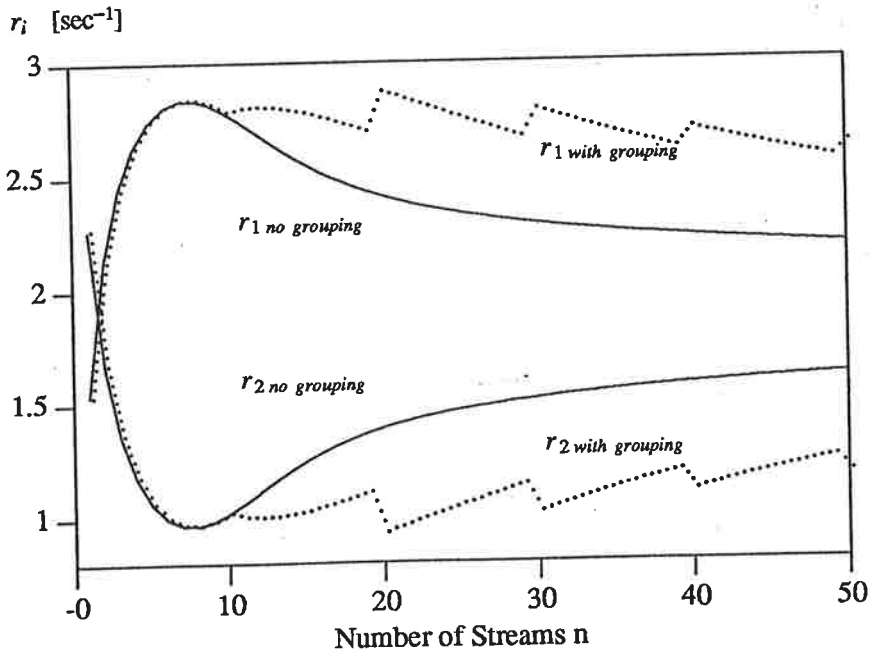


Figure 5.7: Effect of Stream Grouping on Phase Rates r_i Using Method II

streams n increases. On the other hand, Method I is insensitive to stream grouping giving the same results as with superposition carried out using all original component streams.

In order to obtain some comparison of queueing performance displayed when the two methods are used, the superimposed streams were applied to a queue. The two-state MMPP/D/1/N queue was used, as developed in Chapter 4. Figure 5.8 depicts the cell loss probability obtained with the queue size $N = 10$ and three different levels of utilisation ρ. The two-state MMPP stream, used as the input to the queue, was obtained using both Method I and Method II. In the case of Method II superposition was also carried out with and without stream grouping respectively. As can be seen in Figure 5.8, Method I gives most conservative results for the cell loss probability. The two methods vary in their prediction of cell loss probability when the number of the superimposed streams is small. However, as the number of streams increases the cell loss probability values obtained with the two methods converge.

Figure 5.9 shows the corresponding results for cell loss probability when the queue size was increased to $N = 20$.

5.4 Traffic Stream Superposition Algorithm

Since there is is no baseline against which the queueing performance predicted when using the two superposition methods can be compared, it is not clear which method produces more accurate results. In this thesis the choice of one superposition method over the other will therefore be made mainly on the basis of how easily implementable the method is. As a result, Method I will be used in the implementation of the connection admission algorithm. From the results obtained above this method also seems to produce more conservative estimate of cell loss probability. This method is also more general as it does not depend on the assumption that individual streams are modelled as two-state MMPPs.

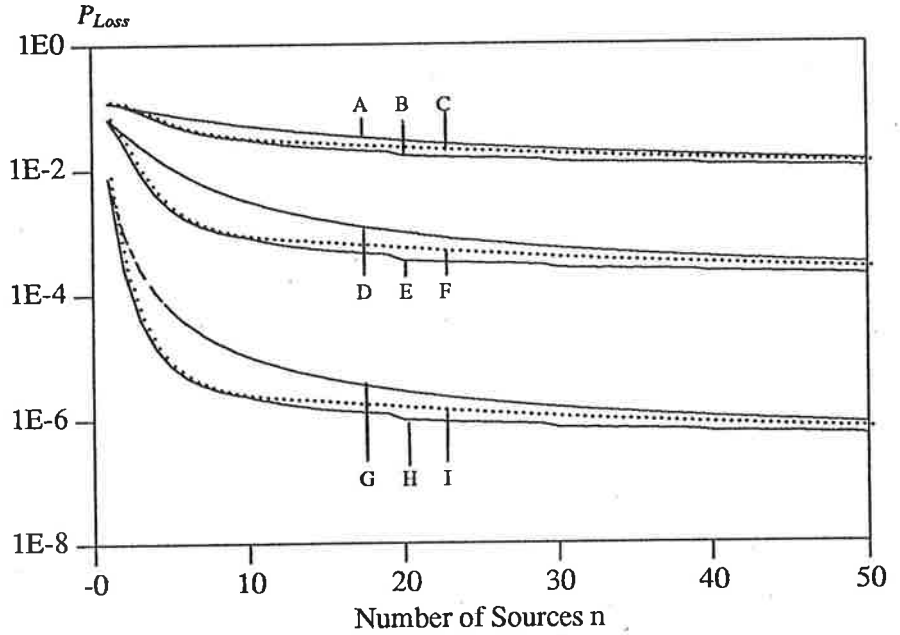


Figure 5.8: Loss Probability for N=10 Using Superposition Methods I and II.

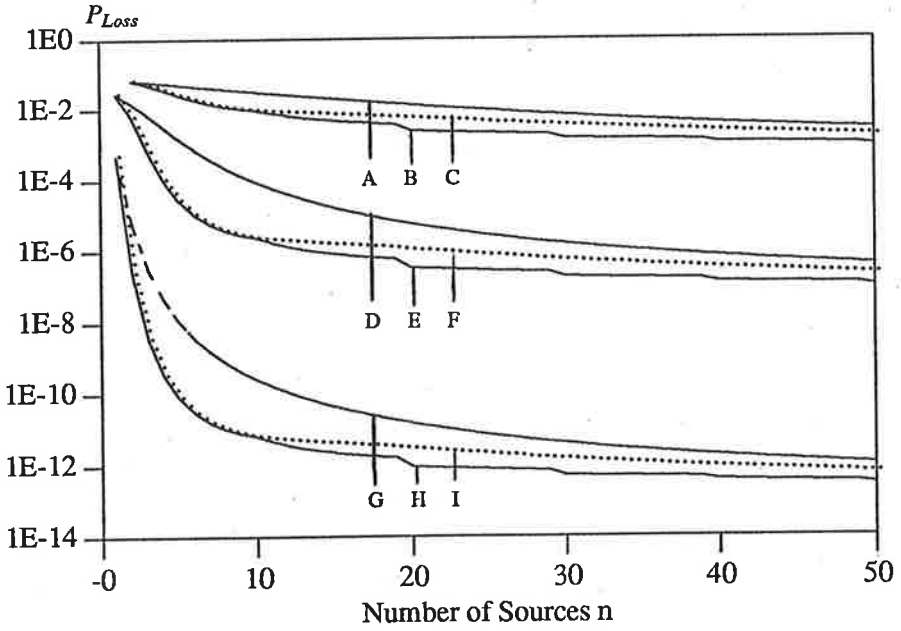


Figure 5.9: Loss Probability for N=20 Using Superposition Methods I and II

A : Method I, $\rho = 0.8$ *B* : Method II with grouping, $\rho = 0.8$ *C* : Method II without grouping, $\rho = 0.8$

D : Method I, $\rho = 0.6$ *E* : Method II with grouping, $\rho = 0.6$ *F* : Method II without grouping, $\rho = 0.6$

G : Method I, $\rho = 0.4$ *H* : Method II with grouping, $\rho = 0.4$ *I* : Method II without grouping, $\rho = 0.4$

The steps which have to be taken in order to obtain a two-state MMPP representation for superposition of n streams using Method I are listed below:

1. Obtain sets of characterising parameters $U_{C_i} = (m_i, v_i, \mu_i^*, \tau_i)$ from sets of defining parameters $U_{D_i} = (\lambda_{1_i}, \lambda_{2_i}, r_{1_i}, r_{2_i})$ of individual two-state MMPPs (Equations 5.1-5.4). This step can be bypassed if the characterising parameters can be obtained directly from the component streams without the two-state MMPP approximation.

2. Obtain the characterising set $U_{C_A} = (m_A, v_A, \mu_A^*, \tau_A)$ for the aggregate stream (Equations 5.5-5.8).

3. Obtain the set of defining parameters $U_{D_A} = (\lambda_{1_A}, \lambda_{2_A}, r_{1_A}, r_{2_A})$ for the two-state MMPP representing the aggregate stream (Equations 5.9-5.12).

The steps required in the above algorithm are also depicted in Figure 5.10.

5.5 Results of Superposition of Heterogeneous Streams

Admission control in future ATM networks will have to deal with a large number of different connections which will be set up to carry different services. In the connection admission model presented in Chapter 3, different services which require the same quality of service from the network may share the same virtual path through the network. As a result, traffic on a virtual path may consist of a large number of dissimilar traffic streams.

In the connection admission model presented in Chapter 3 it was also assumed that cell loss probability (and hence the quality of service) provided in a virtual path was determined by the policing mechanism, which provided peak limiting of the available capacity. This policing mechanism was subsequently modelled as a short queue with a deterministic server. As a result, the performance of the admission mechanism can be studied using such a queue. Taking into account the two-state MMPP model of traffic streams, the queuing system under consideration becomes a

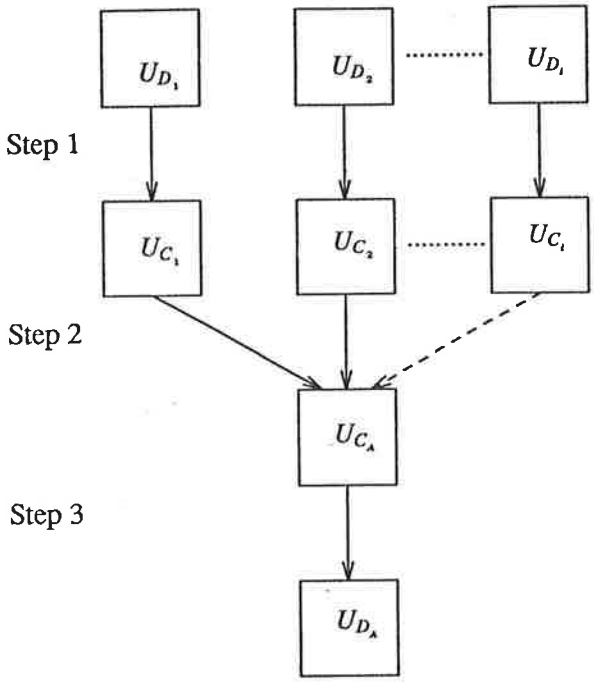


Figure 5.10: Steps Required in Superposition Algorithm

two-state MMPP/D/1/N queue.

Based on this queueing model, an admission algorithm was produced by combining the algorithm for carrying out stream superposition using Method I with the algorithm for solving the two-state MMPP/D/1/N queue, which was presented in Chapter 4. The criterion for stream admission was taken as the overall cell loss probability allowed for the superimposed stream. This algorithm was then used to study admission regions. Linearity of such admission regions may make it feasible to use simpler rules (based, for example, on the concept of the effective bandwidth) for admitting connections in a connection admission algorithm.

Two traffic types representing voice (Type I) and video (Type II) were chosen as input streams into a virtual path with the aim of studying the nature of the admission regions. A voice stream was modelled as an IPP with the following parameters: $\lambda_1 = 54.44$ *cells/sec*, $\lambda_2 = 0$, $r_1 = 2.268\ \text{sec}^{-1}$, $r_2 = 1.532\ \text{sec}^{-1}$. A video stream based on the studies in [39] was modelled also by an IPP with the following parameters: $\lambda_1 = 11.1 * 10^3$ *cells/sec*, $\lambda_2 = 0$, $r_1 = 3.25\ \text{sec}^{-1}$, $r_2 = 0.64\ \text{sec}^{-1}$. The aim of

choosing the above traffic streams was to obtain two representative traffic streams which would reflect the general nature of voice and video connections. They also displayed very different characteristics and hence acted as good examples of dissimilar traffic types which may be expected to be present in future ATM networks.

Having modelled the voice and video streams by IPPs, different mixes of both types were applied to the queue. The set-up is shown pictorially in Figure 5.11. Figures 5.12 to 5.15 show the results obtained with different input stream and queue configurations.

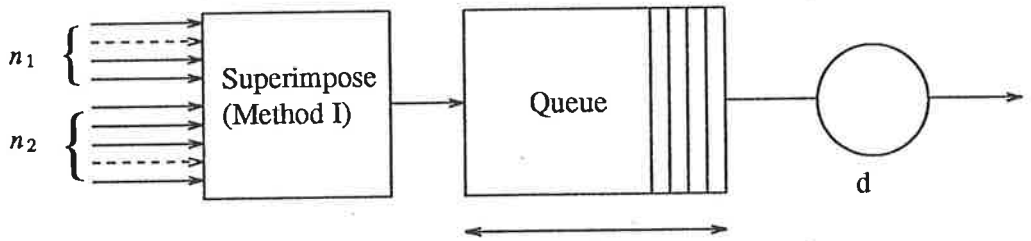


Figure 5.11: Configuration Set-up for Admission Boundary Studies

In Figure 5.12 the service rate d was kept constant at $d = 150$ Mbps and admission boundaries for cell loss probability of $P_L = 1.0\,e^{-8}$ were obtained for queue sizes $N = 5, 7, 10,$ and 20. The near linearity of the resultant admission boundaries suggests that linear approximations may be possible.

In Figure 5.13 the overall utilisation of the queue for the same conditions as those for Figure 5.12 was obtained. It can be seen that decreasing the queue size N results in very low utilisation in order for the required cell loss probability ($P_L = 1.0\,e^{-8}$) to be satisfied. This may have important implications on setting the size of the leaky bucket policing a virtual path at the peak bit rate. Setting $N = 1$, which would ensure complete separation of traffic streams belonging to different virtual paths, might result in an unacceptably low utilisation. This effect may be less pronounced when the discrete nature of the arrival process is taken into account. However, due to jitter in the multiplexing equipment in the network, peak policing may still have

to be relaxed to allow some cell bunching at the peak rate. The choice of the queue size may, as a result, have to be made in a pragmatic fashion. Low utilisation obtained when low cell loss probability (or high quality of service) is required may also have implications on the pricing structure in the future ATM networks.

Figure 5.14 shows admission boundaries for $P_L = 1.0e^{-8}$ with different service rates d while keeping the queue size constant at $N = 10$. Service rates d ranged from $d = 25$ to $d = 150$ Mbps. The linear nature of the admission boundaries can be also observed in this case. However, as the service rate decreases (which corresponds to reducing the capacity available to the virtual path) the boundaries become more convex.

In Figure 5.15 utilisation levels of the queue for the same conditions as in Figure 5.14 are shown. As well as being low, utilisation of the queue decreases more rapidly when the virtual path capacity is low and as the number of Type II (video) streams is increased.

The consequence of near linearity of admission boundaries may be exploited in producing approximate admission algorithms. Boundaries may be approximated by straight lines connecting end points or by straight lines tangent to the admission boundary. Note, however, that approximations based on straight lines connecting end points may lack in accuracy as suggested in [40]. These approximations can be also generalised to higher dimensions as the number of different traffic types increases. The impact of the convexity noted in Figures 5.12 and 5.14 will have to be studied further using more traffic types in order to decide on the best approximation method to be used.

A similar result concerning near linearity of the acceptance region was noted in [20] and [3] where a bufferless network was assumed, with traffic modelled as interrupted deterministic streams, and where the statistical behaviour was studied using a large deviation approximation.

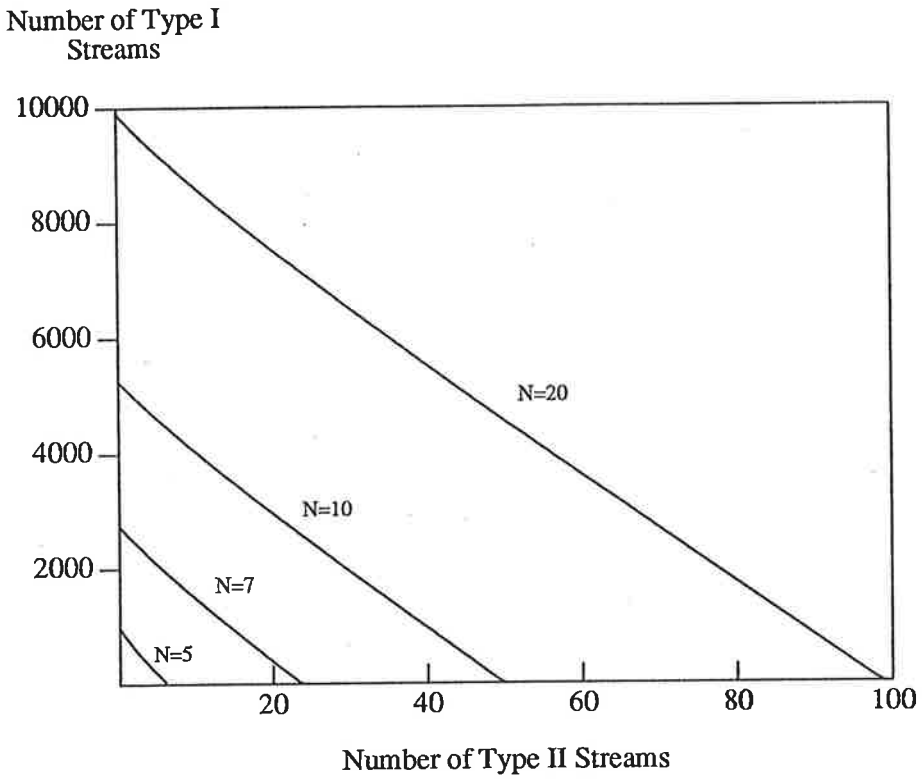


Figure 5.12: Admission Regions for Different Queue Sizes N

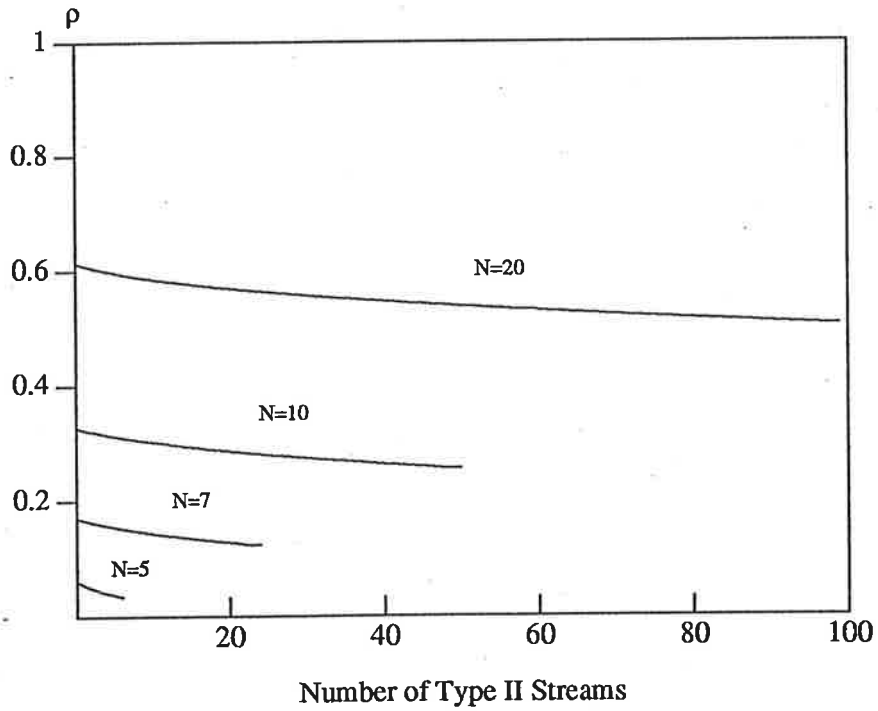


Figure 5.13: Effect of Traffic Mix and Queue Size on Utilisation

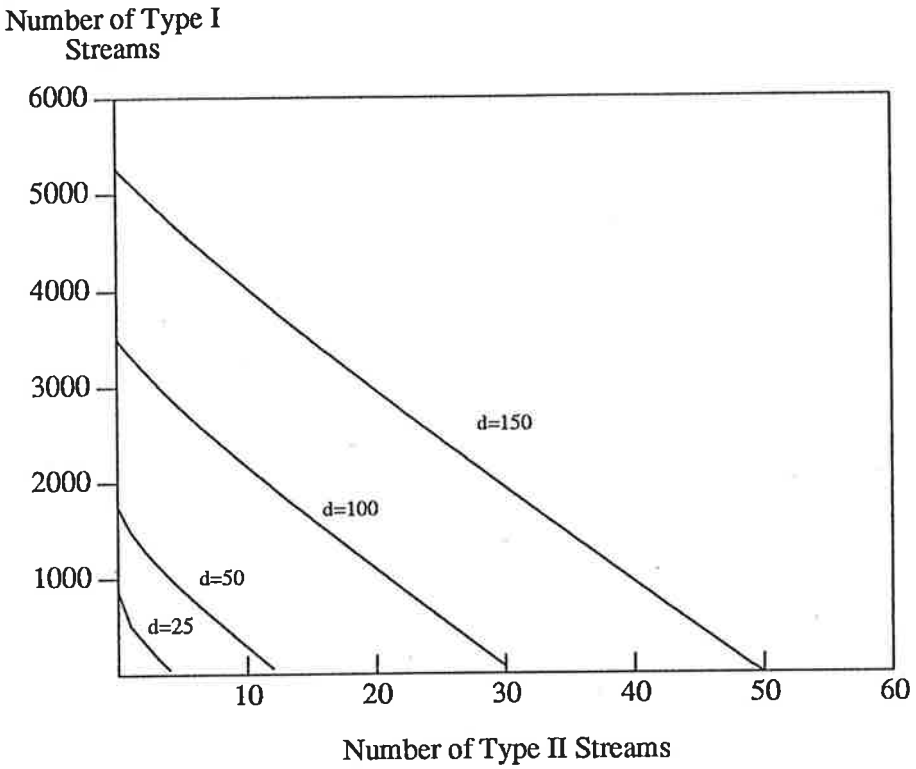


Figure 5.14: Admission Regions for Different Queue Service Rates d

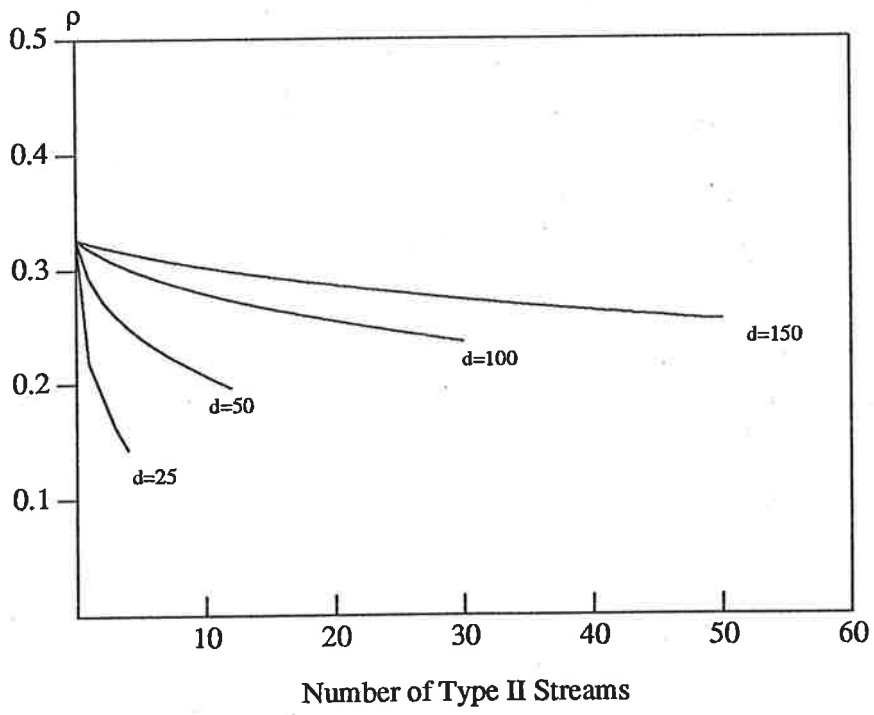


Figure 5.15: Effect of Service Rates d and Traffic Mix on Utilisation

In Figure 5.16 the results of a study aiming at determining the sensitivity of the required effective bandwidth as cell loss probability in the virtual path changed, are shown. Type I streams only were used in this case. Effective bandwidth was defined as the available capacity over the maximum number of streams carried for a particular value of N and the required cell loss probability P_{Loss}. (This differs from the studies carried out in [20] where the effective bandwidth depended on the available bandwidth and the required cell loss probability.) The available capacity of the virtual path was set at $d = 150$ Mbps. The curves were obtained for a number of queue sizes N.

It can be observed from Figure 5.16 that as the queue size N was increased, changes in many orders of magnitude to cell loss probability resulted in very small changes to the effective bandwidth requirement. On the other hand, it should be noted that only a small change in effective bandwidth may result in changes to cell loss probability of many magnitudes. For example, for $N = 20$ and $P_{Loss} = 1E{-}8$, 5% change in the value of the effective bandwidth results in changes to P_{Loss} by two orders of magnitude. This change becomes smaller as N is decreased. This is, however, a much less pronounced effect than that observed in [20] where a 5% change in the value of the effective bandwidth resulted in changes to cell loss probability by eight orders of magnitude. Such dependence could have important consequences on the sensitivity of the quality of service provided in a virtual path.

5.6 Conclusion

Approximation methods should be able to characterise real life traffic in terms of its effect on the queueing system to which the traffic is applied. Consequently, any superposition method should result in the analysis which faithfully reflects the impact of the superimposed stream on that queueing system. Traffic characterisation which will allow for accurate queueing performance prediction, however, is still a topic of some debate. Further research must be carried out to provide solutions to this problem.

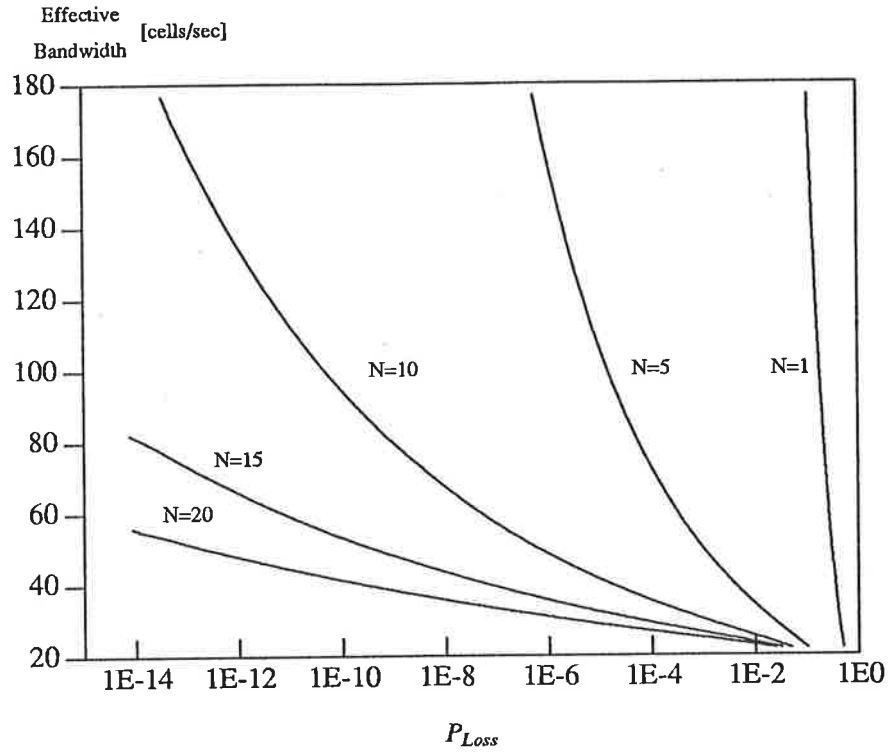


Figure 5.16: Effect of Effective Bandwidth on Loss Probability

Bearing the above observations in mind, two superposition methods were presented and compared in this chapter. The aim of the comparison was their suitability for implementation in a connection admission algorithm. Based largely on their ease of implementation, Method I, first proposed in [24], was chosen. This method is also more general and does not depend on the assumption that individual traffic streams are modelled as two-state MMPPs, although it was implemented by modelling individual traffic streams in this manner.

Using superposition Method I together with the algorithm for the two-state MMPP/D/1/N queueing system modelling a virtual path, superposition of heterogeneous traffic streams was studied. Admission boundaries displaying linear nature were observed. This linearity could be exploited in developing connection admission algorithms based on the concept of the effective bandwidth. Low utilisation was also observed when small values of queue length N were used and when the available capacity was low compared to the capacity requirements of

individual connections. This may affect the choice of the depth of the leaky bucket policing the virtual path and the capacity allocated to each virtual path. It was also noted that the cell loss probability experienced by connections in a virtual path was highly dependent on the amount of capacity allocated to these connections. This fact could also have important consequences on the sensitivity of the quality of service provided in a virtual path.

In Chapter 6 the details of the connection admission algorithm, making use of superposition Method I and algorithmic solution of the two-state MMPP/D/1/N queue (given in Section 4.4 of Chapter 4), are presented. This connection admission algorithm can be used for more extensive studies of the nature of admission boundaries with a wide range of different traffic types and virtual path conditions.

Chapter Six:

Connection Admission Control Algorithm

6.1 Introduction

In the previous chapters the model for connection admission control was developed and algorithmic analysis for the model was presented. The model was analysed in two parts. Part one presented in Chapter 4 dealt with the analysis of the queueing model for the virtual path. A detailed algorithm for the solution of the resultant two-state MMPP/D/1/N queue was also included in that chapter. Part two of the connection admission control model involved superposition of traffic streams modelling individual connections in the network in order to obtain the two-state MMPP representation of the arrival stream. In Chapter 5 studies of two superposition methods were carried out allowing superposition of two-state MMPP streams, and one particular method was proposed for use in the connection admission algorithm. Further studies of connection admission regions were carried out in Chapter 5 using heterogeneous traffic streams. Near linearity of admission boundaries was noted: a result which could be exploited in developing simple connection admission algorithms based on the concept of the effective bandwidth.

In this chapter the above studies are brought together and applied in developing an algorithm for connection admission control. The algorithm makes use of the virtual network concept described in Chapter 3 and it incorporates algorithms developed in Chapters 4 and 5. The aim of this algorithm is to act as a reference from which

simpler connection admission algorithms can be developed and which can be used for verification and comparison with other similar algorithms. The algorithm can also be used for further studies of the nature of admission regions with a large variety of traffic types. Performance analysis of this connection admission algorithm, however, is not presented in this dissertation.

A general description of the algorithm is presented in Section 6.2 followed by a detailed description of the steps involved. These steps are then summarised in the subsequent flow chart.

6.2 The Algorithm

There are two distinct actions involved in connection admission control. These involve respectively addition of a new connection and removal of an old connection. Figures 6.1 and 6.2 describe general steps required in carrying out these functions. The algorithm presented below prescribes the individual steps to be taken.

Algorithm Steps

Adding New Connections

1. Traffic parameters describing the expected traffic volume of a connection presented in the call set-up request must be converted into a form suitable for carrying out superposition with traffic streams of other connections sharing the same virtual path. Although no final recommendation has been made on the form these parameters should take, [9] recommends parameters such as: mean bit rate, peak bit rate, and average burst duration. If individual connections are modelled as two-state MMPPs or if the superposition is to be modelled as a two-state MMPP, a fourth parameter is also required for complete specification.

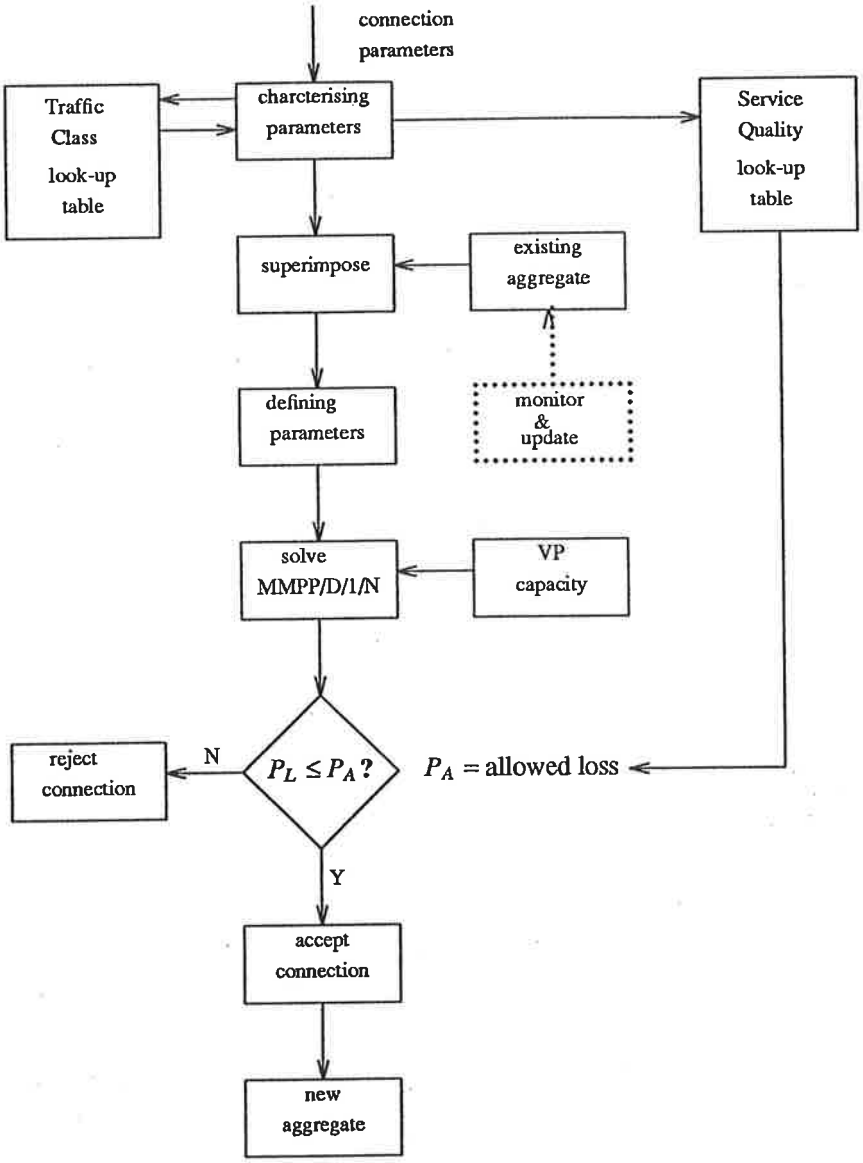


Figure 6.1: Addition of New Connections

Considering the above, the following connection parameters are suggested in order to describe traffic volume to be carried in a connection: mean bit rate (*mean*), peak bit rate (*peak*), average duration of high activity period T_A, and average duration of low activity period T_I. The above choice is influenced by modelling individual connections by two-state MMPPs. In fact, high and low activity periods correspond to durations of state 1 and state 2 of the two-state

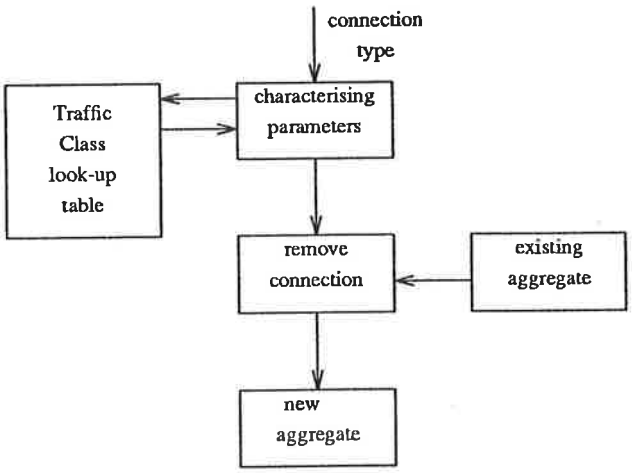


Figure 6.2: Removal of Old Connections

MMPP model as presented in Chapter 4.

Connection parameters are related to the defining parameters of a two-state MMPP in the following manner:

$$mean = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}$$

$$peak = \lambda_1$$

$$T_A = \frac{1}{r_1}$$

$$T_I = \frac{1}{r_2}$$

The four parameters which were used for traffic characterisation in superposition Method I in Chapter 5, (i.e. mean m, variance v, third moment μ^*, and time constant τ) can then be obtained from the above connection parameters using Equations 5.1-5.4 in the following manner:

$$m = mean,$$

$$v = \frac{T_A}{T_I}(peak - mean)^2,$$

$$\mu^* = \frac{T_A peak^3 + T_I\,((1 + \frac{T_A}{T_I})\,mean - (\frac{T_A}{T_I})peak)^3}{T_A + T_I}$$

$$- 3\,\frac{T_A}{T_I} mean\ (peak - mean)^2 - mean^3,$$

$$\tau = \frac{T_A\, T_I}{T_A + T_I}.$$

In the above conversion the peak bit rate was made equal to the mean bit rate in state 1 of the two-state MMPP. This is a conservative choice and other methods of matching the peak rate of the connection to the parameters of the two-state MMPP could be considered.

Not all traffic types need all four parameters to describe their characteristics. In general, three separate traffic type cases may be distinguished:

i. Continuous bit rate traffic and traffic which can be represented by a Poisson approximation.
   One parameter m suffices in this case. Such traffic types can be specified explicitly or implicitly by setting: $peak = mean$.

ii. Traffic which can be represented by an interrupted Poisson process.
   Three parameters are required in this case. These are: *mean*, *peak*, and T_A.

iii. Traffic which is to be approximated by a two-state MMPP.
   All four parameters: *mean*, *peak*, T_A, and T_I are needed.

2. In order to determine the characterising parameters from the connection parameters, the above conversion is necessary. However, in order to able to carry out this conversion off-line and to avoid storing the parameters for each connection for the duration of the connection, traffic classes corresponding to traffic types recognizable by the network can be used. A traffic class can be associated with a set of predefined parameters. The traffic class for a given connection can then be determined by comparing the set of connection parameters of the given connection with the available sets of predefined parameters using look-up tables. The set of predefined parameters which gives the closest match to the set of connection parameters will determine the

traffic class for that connection. Characterising parameters based on this set of predefined parameters are then used in the remaining steps of the algorithm to describe the connection. A traffic class indicator can also be associated with the connection in order to determine the characterising parameters needed when the connection is to be removed.

Note also that the conversion of connection parameters to characterising parameters based on the two-state MMPP model of the connection could be avoided if characterising parameters for the connection were specified directly by the user. They may, however, be difficult to estimate and measure.

3. To determine the virtual path required to carry the connection, the route for the connection and the quality of service demanded from the connection are required. In this algorithm only the quality of service parameter is used. This parameter is required so that the cell loss probability measure for the virtual path can be obtained. As was the case with traffic classes, a number of predefined quality of service classes can also be provided in the network. A connection will then be assigned a virtual path which provides quality of service matching most accurately that demanded by the connection.

   As a result of the above, a virtual path corresponding to a particular cell loss probability P_{Loss} will be selected.

   In a similar fashion grade of service classes (corresponding to connection blocking probability) could be provided. An additional connection parameter specifying the required grade of service for the connection would have to be supplied. To take into account the grade of service requirement the admission algorithm may, however, have to be altered.

4. The current state of each virtual path must be known at all times. This can be achieved by associating with each virtual path traffic parameters describing the aggregate traffic stream in that virtual path. These traffic parameters

could also be updated periodically by monitoring traffic present in the virtual path.

In order to determine the state of the virtual path with the new connection accepted, the existing aggregate traffic (E) characterised by (m_E, ν_E, μ_E^*, τ_E) must be combined with the traffic corresponding to the new connection (N) described by (m_N, ν_N, μ_N^*, ,τ_N). The parameters characterising the 2-state MMPP describing the new aggregate (NA) can be obtained from:

$$m_{NA} = m_E + m_N,$$

$$\nu_{NA} = \nu_E + \nu_N,$$

$$\mu_{NA}^* = \mu_E^* + \mu_N^*,$$

$$\tau_{NA} = \frac{1}{\nu_{NA}}(\tau_E \nu_E + \tau_N \nu_N).$$

These characterising parameters need to be subsequently inverted in order to obtain the defining parameters of the two-state MMPP ($\lambda_{1_{NA}}, \lambda_{2_{NA}}, r_{1_{NA}}$, and $r_{2_{NA}}$) for the new aggregate traffic stream. The details of the inversion were presented in Chapter 5 (Equations 5.9-5.12).

5. The defining parameters of the two-state MMPP for the new aggregate traffic stream ($\lambda_{1_{NA}}, \lambda_{2_{NA}}, r_{1_{NA}}$, and $r_{2_{NA}}$) are then used as input into the queueing system modelling the virtual path. The algorithm for solving this system was presented in Section 4.4 of Chapter 4. The other parameters required for the solution are the service rate of the queue d, and the length of the queue N. Since the purpose of the queueing system is to limit the peak capacity available to the virtual path, the service rate d corresponds to the capacity of the virtual path and the length of the queue should be set to a small constant value which may have to be chosen in a pragmatic way as discussed in Chapter 5.

The output of the queueing system algorithm is the cell loss probability P_{Loss_N} which would result with the new connection accepted.

6. If

$$P_{Loss_N} \leq P_{Loss},$$

then the new connection can be accepted, otherwise it should be rejected. If the new connection is accepted, the current state of the virtual path needs to be updated by saving the parameters of the new aggregate traffic present in the virtual path.

Removing an old connection

1. In order to remove an old connection, traffic parameters describing this connection must be known. If traffic classes are used by the network, the knowledge of the traffic class the connection belongs to can be used to obtain the required parameters using an appropriate look-up table. Using traffic classes avoids having to save traffic parameters for every connection. This would require a large amount of memory space which might have to be managed very fast in order to keep up with the state of the network.

   The required characterising parameters for the connection to be removed (R) are (m_R, ν_R, μ_R^*, τ_R). The characterising parameters for the new aggregate traffic after the removal of the connection can be obtained from:

$$m_{NA} = m_E - m_R,$$

$$\nu_{NA} = \nu_E - \nu_R,$$

$$\mu_{NA}^* = \mu_E^* - \mu_R^*,$$

$$\tau_{NA} = \frac{1}{\nu_{NA}}(\nu_E \tau_E - \nu_R \tau_R).$$

2. The parameters of the new aggregate traffic in the virtual path need to be saved. They will be used to determine the current state of the virtual path when new connections are added or old connections are removed from the

virtual path. These parameters may also be periodically updated after measuring the traffic present in the virtual path.

Figure 6.3 shows a summary flowchart for the process of adding and removing connections as described above.

The connection admission algorithm described in this chapter, as well as acting as a reference algorithm for further studies into simpler connection admission algorithms, could also be used in practice in an ATM network. However, certain steps employed in the algorithm might have to be simplified. In particular, the step involving the algorithmic solution of the two-state MMPP/D/1/N queueing system may prove computationally too expensive for practical implementation. In this case methods based, for example, on fluid flow approximations [49] might be used to obtain the required solution for the queue.

6.3 Conclusion

A connection admission algorithm has been presented based on the connection admission model proposed in Chapter 3. The algorithm prescribes the actions to be taken for addition and removal of connections from an ATM network. It incorporates the two algorithms developed for the two parts of the connection admission model as presented in Chapters 4 and 5.

The main purpose of this algorithm is to act as a reference model from which other simpler connection admission algorithms can be developed and which can be used for comparison with other similar algorithms. In particular, this algorithm could be used to study the nature of admission boundaries produced when dissimilar connections are combined. This, in turn, could lead to connection admission algorithms based on the concept of the effective bandwidth.

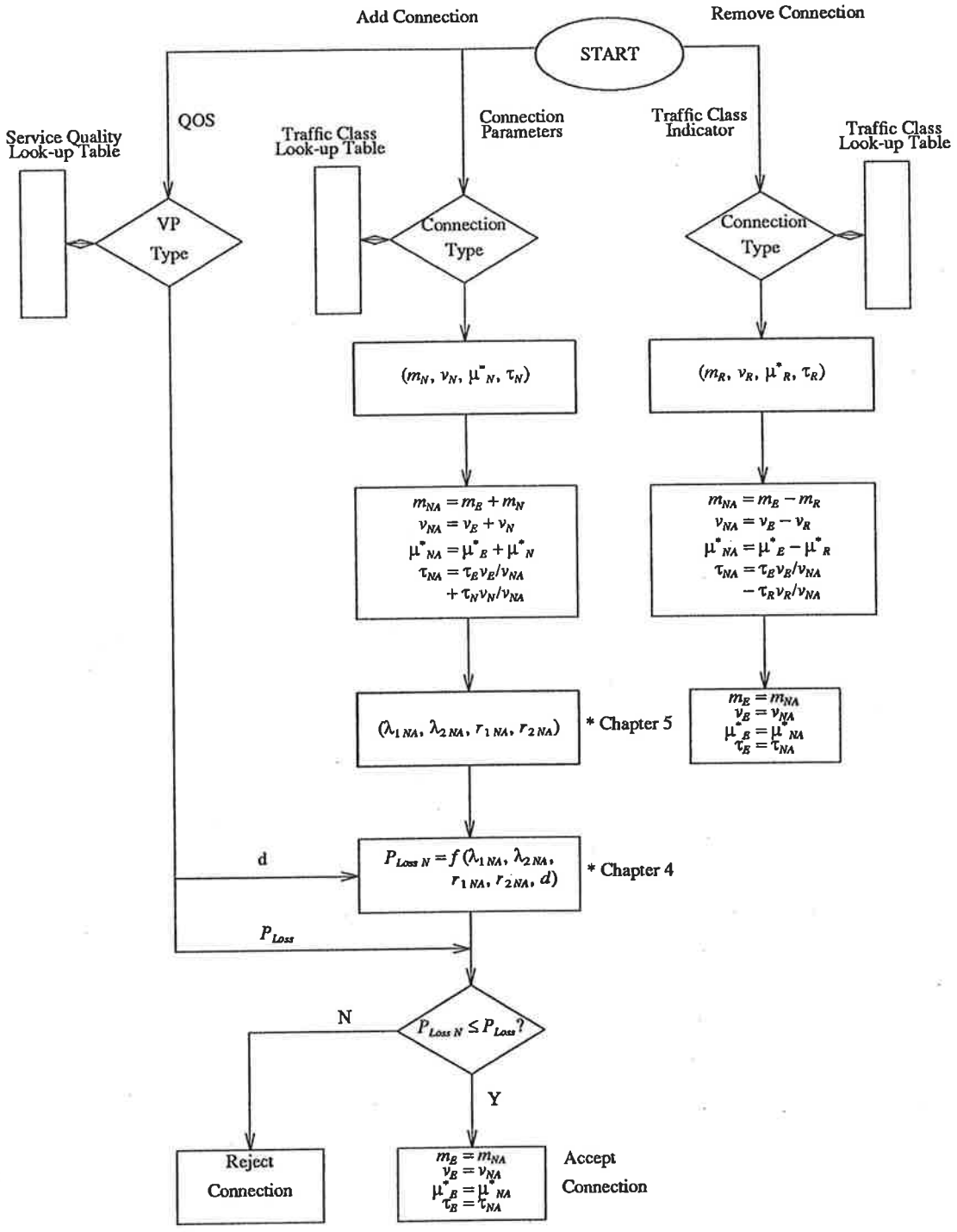


Figure 6.3: Flow Chart for Connection Admission Control

Chapter Seven:

Conclusion

Connection admission control forms an important aspect of traffic control in ATM networks. Its purpose is to admit or reject new traffic flows based on the required quality of service and the available network resources. The problem of connection admission control becomes quite complex, since a large number of different traffic types and distinct qualities of service must be provided for. However, resultant connection admission control algorithms should be simple and efficient, allowing admission to take place in real time. Approaches proposed in the literature range from those based on queueing models to those based on the concept of the effective bandwidth. The latter approach holds a lot of promise towards achieving simple real time algorithms for connection admission control.

In Chapter 2 a literature survey was carried out to determine the main approaches and methods proposed for resource management and traffic control in ATM networks. A need for an architectural framework was identified, which would allow provision of guaranteed levels of performance in an efficient and flexible fashion.

This architecture, termed traffic management and control architecture, was presented in Chapter 3. It was based on the concept of virtual networks. This approach for provision of distinct qualities of service had several advantages over other methods, which included hardware partitioning of the network and use of explicit priority levels. Connection admission control was based on this new architecture. The connection admission control model proposed in Chapter 3 was subdivided into two parts involving respectively the queueing model for the virtual

path and the stream arrival model for stream superposition.

The queueing model, represented by the two-state MMPP/D/1/N queue, was solved in Chapter 4 for cell loss probability indicating the quality of service in a virtual path. The solution was presented in an algorithmic form, which was also implemented in software.

The stream arrival model for stream superposition was investigated in Chapter 5. Two superposition methods were studied and compared. Due to the lack of any baseline for queueing performance evaluation of the two methods, the choice for using one of them in a connection admission control algorithm was made based mainly on its ease of implementation.

The nature of connection admission boundaries was also investigated in Chapter 5, using heterogeneous traffic streams representing voice and video. It was found that such boundaries displayed near linearity. This was significant as it allowed the possibility of developing simple connection admission control algorithms based on the concept of the effective bandwidth. However, it was also observed that these boundaries exhibited some convexity. This might make linear approximations more difficult. Approximations based on lines tangent to the admission boundaries rather than on lines connecting the end points may have to be investigated.

The connection admission algorithm was presented in Chapter 6. It combined the algorithm for the queueing model of the virtual path given in Chapter 4 and the algorithm for stream superposition presented in Chapter 5. The use of traffic classes and quality of service classes was proposed in this algorithm. Inclusion of grade of service classes for connections was also possible. The possible implementation of this algorithm in real time will depend on its real time performance. However, the algorithm can also be used extensively in studying connection admission boundaries using a large variety of traffic types and virtual path conditions.

The main contentious issue identified in this thesis was the problem of traffic characterisation. It is still very difficult to state which characteristics in a traffic stream should be used to determine its queueing performance. This problem manifested itself in the study of superposition methods and in the choice of traffic parameters in the connection admission control algorithm.

Another problem which will need further study concerns the depth of the leaky bucket policing mechanism used for peak bit rate policing of virtual paths. It was found that the theoretical setting of this depth produced very low utilisations. This problem could be partly an artifact of the continuous time nature of the stream arrival model. However, the final choice of the leaky bucket depth may ultimately have to be made in a pragmatic fashion.

Further work will be required to study the problem of traffic characterisation. The proposed connection admission control algorithm will also have to be analysed to determine its performance and possible implementation as a real-time algorithm in a future ATM network. Approximations, which could improve its timing performance, may also have to be developed. Studies of the nature of connection admission boundaries will also be carried out using this algorithm. The aim of these studies will be development of other simple algorithms based on the idea of the effective bandwidth.

Appendix A:

Traffic Management and Control in Broadband Networks

Paper presented at the 5th Australian Fast Packet Switching Workshop in Melbourne, 9-11 July 1990.

Traffic Management and Control in Broadband Networks

E. Dutkiewicz and G. Anido

Development Unit
OTC Limited
231 Elizabeth St
Sydney NSW 2000
New South Wales, AUSTRALIA

Phone: (02) 2874348
Fax: (02) 2874990
Acsnet: eric@otc.oz

Abstract

A flexible three level management architecture provides a convenient basis for managing and controlling resources in broadband networks. The control and management mechanisms use the concepts of virtual channels, virtual paths, and virtual networks. A simple algorithm using a 2-state Markov modulated approximation for traffic streams can be used for call admission.

1. INTRODUCTION

Broadband networks will be required to support a variety of services displaying a wide range of characteristics and demanding from the network different levels of performance. Of particular interest, at least in the early stages of the existence of broadband networks, will be the problem of supporting existing services which often require deterministic channels with stringent temporal and spatial requirements. Provision will also have to be made for other future services which may place different demands on the network. In order to deal with these various requirements traffic control and traffic management capabilities will have to be provided in such networks.

Of particular interest in the design of management and control strategies in broadband networks is the problem of connection admission control. Its function is to ensure that calls are accepted provided there is enough resources to support them with guaranteed quality without degrading the quality of existing connections. The problem is particularly complex when calls displaying vastly different characteristics must be accommodated.

In this paper our attention is focused on broadband networks which use asynchronous transfer mode (ATM) as their transport mechanisms. Section 2 of the paper gives an overview of the basic concepts of an ATM network which are used in resource management and control strategies. In Section 3 the different levels of the management architecture are discussed. Possible admission control strategies including a simple call admission control algorithm based on a two-state Markov-modulated model of traffic streams are presented in Section 4.

2. ATM NETWORK OVERVIEW

The basic communication channel provided in an ATM network will be the virtual channel. Each virtual channel will be unidirectional and it will support information transfer between two or more endpoints in the network. Each virtual channel will carry ATM cells belonging to a single call [3].

Virtual channels can be grouped in the network to form virtual paths. Virtual paths are also unidirectional interconnecting a given pair of exchanges in the network. The use of virtual paths in ATM networks will allow for simplification of the routing and control mechanisms in the core network. This will result in reduction of call processing time and cost. Fast recovery from network equipment failures will also be facilitated through fast re-routing capability without the involvement of individual

virtual channels.

In order to distinguish between different qualities of service, which might be demanded from the network, each virtual path can be associated with a particular quality of service which leads to the idea of virtual networks or logical overlay networks [1]. A virtual network consists of virtual paths which provide the same or very similar quality of service to network users.

3. MANAGEMENT ARCHITECTURE

The management architecture is required which will provide a convenient basis for managing and controlling ATM network resources. The aim is to be able to accommodate a wide range of services and to make efficient use of the available resources. In building the management architecture we utilise the concepts of virtual channels, virtual paths, and virtual networks which provide communication channels within ATM networks.

The three levels of the management architecture include the ATM Resource Management level, ATM Traffic Management level, and ATM Traffic Control level. The ATM Resource Management level creates virtual networks and manages available transmission bandwidth and other network resources which are made available to virtual networks. The ATM Traffic Management level is concerned with the management of available bandwidth within virtual paths, whereas the ATM Traffic Control level deals with traffic flows within virtual paths and virtual channels. Figure 1 shows the various levels in the management architecture and the manner in which they relate to virtual channels and virtual paths.

3.1 ATM Resource Management

The ATM Resource Management level is responsible for setting up and managing virtual networks as demand for them changes in the network. This level monitors the use of resources in virtual paths and adjusts the allocated resources as demand falls or rises.

3.2 ATM Traffic Management

The ATM Traffic Management level is responsible for managing capacity in individual virtual paths. At this level decisions are also made whether a call should be accepted into the network. Admission decisions will have to be based on traffic parameters describing the traffic flow in calls, the quality of service and the grade of service demanded by calls, and the resources available in the network. Service classes may also be defined based on the grade of service which they require. Grade of service may in turn be related to the tariffing structure.

3.3 ATM Traffic Control

The ATM Traffic Control level is responsible for ensuring that traffic flows in individual virtual channels and virtual paths do not exceed allowed thresholds. Due to the high speed of information flow in ATM networks preventive control rather than reactive control will have to be exercised [14]. The main purpose of this usage monitoring mechanism at the virtual channel level is to protect network resources from other users who by exceeding their negotiated traffic parameters might otherwise degrade the quality of service of other users. The simplest, and safest, action that can be taken upon detecting traffic violation is cell discard in the offending stream.

At the virtual path level the aim of the traffic control level is to maintain separation between different virtual path so that any interference between cells belonging to different virtual paths is minimised. This action will ensure that qualities of service associated with different virtual paths are protected. A simple way of implementing this separation can be achieved by policing each virtual path at peak rates.

4. CALL ADMISSION

Call admission into an ATM network will be based on satisfying two criteria: quality of service and grade of service. Quality of service is a performance measure used at the ATM cell level whereas grade of service is used at the call level. Quality of service is often associated with cell loss probability and grade of service with call blocking probability. Different grades of service may be related to different classes of service. For example, two calls belonging to different classes of service which may require admission into the same virtual path may be treated differently as a result of their different grades of service.

The approaches taken in the literature concentrate on the quality of service aspects of the call admission control. The common approach relies on the knowledge of the call bit rate distributions. The statistical parameters used to describe such distributions include average bit rate, bit rate variance , and peak bit rate. In [12] average bit rate and bit rate variance of individual calls is

used in the admission algorithm. It is assumed that when a sufficient number of such calls is multiplexed together the resulting bit rate distribution will approach a normal distribution. This will then allow the determination of the maximum allowable load which can be permitted for a given cell loss performance.

A similar approach is taken in [13]. In this case the acceptance decision is based on the mean bit rates and peak bit reates of the new and existing calls without the need for the bit rate variance.

The approaches based on the knowledge of bit rate distributions do not take into account the fluctuations of bit rates in time. Thus a model which incorporates a measure of burstiness might be more useful.

In our proposed management architecture the function of call admission will be performed at the ATM Traffic Management level. A simple call admission scenario which does not take into account different grades of service can look as follows: A call is requested with a particular set of parameters providing information about its source characteristics and demanding a particular quality of service. An appropriate virtual network which will provide the required quality of service must be selected, within that virtual network a virtual path to the required destination must be found. To determine whether the call should be admitted into the virtual path its affect on the quality of service of calls already supported in that virtual path must be determined. If the quality of service is not degraded below the level required in the virtual path then the call can be admitted. In case of call rejection, provision may also be made for negotiation between the user and the network in which a new set of source and performance characteristics should be agreed upon.

Admission of calls into the network should not, however, be based entirely on the current state of the traffic in the virtual path. It is possible to envisage situations where due to a wide range of service classes vying for admission some calls should be rejected even though there might be enough capacity to satisfy the cell loss performance of the virtual path at the time when the acceptance is required. This would be especially significant if distinct grades of service were associated with different service classes (resulting, for example, from different tariffing levels for different service classes). As a result, it could be more advantageous to reject a low reward call if there is a high probability of a high reward call being requested in the future.

Different grades of service associated with different service classes may also result in accepting certain calls even though the resulting quality of service produced on the virtual path at the time of admission indicates that such a call should be rejected. Such a decision might be based on the expected reward and again requires the knowledge of the future state of the virtual path.

The discussion presented above indicates that call admission may have to be based not only on the quality of service criterion but also on the grade of service required by different service classes. Satisfying quality of service would be adequate only if grades of service associated with all service classes were identical.

Quality of service is associated with the cell level performance on the virtual path whereas grade of service is associated with call level performance. Our approach is to study call admission based on the quality of service measure as a first step and then apply the grade of service measure as a modifying condition for the final call acceptance decision.

4.1 Call Admission Model

In this model we assume that the grades of service required by all calls requesting admission into the virtual path are identical. This allows us to concentrate on the quality of service measure as a criterion for admission.

The quality of service measure can be defined in terms of cell loss performance and cell delay performance. Cell delay performance will, to a large extent, depend on the transmission delays which in turn depend on the size of the network and the route taken through the network. Buffers in the network will not be large because their implementation would be too complex and they would penalise cell delay and cell jitter of delay sensitive services [6]. As a result, it seems more important to associate quality of service with cell loss. The important question now is whether it is sufficient to describe cell loss in terms of long-term average cell loss or whether other measures such as frequency of lossy periods and duration of lossy periods should be also considered [15].

As mentioned above, in order to guarantee different qualities of service in the network without using explicit priorities, virtual paths can be utilised. Each virtual path will, as a result, have

to support a given quality of service. In order to ensure that distinct qualities of service are provided virtual paths will have to be separated in the network. Such separation can be accomplished by peak limiting the capacity utilised by each virtual path. The policing mechanisms will be placed in the network at places wherever virtual paths originate. The quality of service will then be determined by the policing mechanism and call admission mechanism.

4.1.1 Virtual Path Model

One policing mechanism that has received a lot of attention in the literature is based on the leaky bucket algorithm [2]. In cases where only peak cell rates are to be policed such a mechanism can be modelled as a finite queue with a single deterministic server. (G/D/1/N in Kendall notation) where G denotes a general arrival process, the service rate corresponds to the allowed peak rate and N = 1. Such a set-up assumes that no 'bunching' of cells is allowed. A more realistic approach would be to adjust N allowing a short burst of cells to pass through. However, in order to maintain separation between virtual paths N should be kept small.

The quality of service of the calls carried by the virtual path will depend on the cell loss performance displayed by the virtual path. In order to be able to control this performance measure for different virtual paths, cell loss in the core network will have to be very low so that it is the policing mechanisms which affect cell loss suffered by cells in individual virtual paths.

Taking the above into account, our model of the virtual path can be reduced to the model of the policing mechanism which determines the cell loss performance and hence the quality of service provided by the virtual path.

4.1.2 Cell Arrival Model

In order to determine the cell loss performance of the virtual path we have to model cell arrival processes of individual calls. A useful summary of arrival processes studied in the literature for voice, video, and data services is presented in [7]. Our aim is to find a model which can be used to approximate a large number of different call types. The canonical process which we choose for this purpose is the two-state Markov-modulated Poisson process (MMPP). This process has been used in the literature to approximate much more complicated processes very accurately. (see, for example [5]). Special cases of the two-state MMPP include the Poisson process and the interrupted Poisson process (IPP) which find extensive use in telecommunication modelling.

A two-state MMPP can be fully described by four parameters (λ_1,λ_2,r_1, and r_2). The process spends exponentially distributed times with means $1/r_1$ and $1/r_2$ in states 1 and 2 respectively. When the process is in state 1 the arrival process is a Poisson process with intensity λ_1 and in state 2 it is a Poisson process with intensity λ_2. A number of different set of characterising parameters for modelling a traffic stream have been proposed to match different characteristics of the stream [4],[5],[10].

We can use this process to model the aggregate process produced by calls in the virtual path. As a result the model of the virtual path can be represented as a MMPP/D/1/N queue for which solutions can be obtained (see, for example [8],[9],[11]). Moreover, if we model each incoming call as a two-state MMPP, which will allow for a wide range of call types to be accommodated, we will be able to produce a very simple cell level call admission algorithm.

4.2 Cell Level Call Admission Algorithm

Assume that a virtual path has been set up between an origin and destination to provide a quality of service in terms of average cell loss probability P_l. We approximate every incoming call by a two-state MMPP. The aggregate traffic produced by all calls carried in the virtual path (called here the aggregate call) is also approximated by a two-state MMPP.

The decision whether a new call should be accepted or rejected will then be based on determining a new aggregate call including the new call and the existing calls and applying the aggregate traffic stream to the MMPP/1/D/N queue in order to solve for the cell loss probability P_{l_n}. The new call will be accepted if $P_{l_n} \leq P_l$.

The major steps involved in the algorithm can be summarised as follows:

1. Model the incoming call as a two-state MMPP.
2. Obtain the aggregate call by superimposing the new call to the existing calls.
3. Model the aggregate call by a two-state MMPP.

4. Apply the aggregate call to the MMPP/D/1/N queue to obtain the resulting cell loss probability.
5. Accept the new call if the resulting cell loss probability does not exceed the allowed cell loss probability for the virtual path.

Figure 2 attempts to clarify the main steps of the above algorithm.

Different options which may simplify the above algorithm will be studied. These will include various methods for achieving superposition. (Some methods based on different sets of characterising parameters have been put forward in [4],[5], and [10].) Moreover, other schemes, such as maintaining a 'running' superposition which is updated with the arrival of each new call or a number of calls, will also be investigated.

5. CONCLUSION

In this paper we have presented a flexible three level management architecture which can be used for resource management and control in ATM networks. The approach taken allows for provision of distinct qualities of service in the network. Protection of calls requiring different qualities of service is achieved by separating virtual paths which are associated with distinct qualities of service. Finally, a simple call admission algorithm based on qualities of service required by calls and using a two-state Markov-modulated approximation for traffic streams has been proposed. Additional requirements of providing distinct grades of service to different service classes can also be incorporated into the algorithm.

6. ACKNOWLEDGEMENT

The permission of the Managing Director of OTC to publish this work is gratefully acknowledged. The views and opinions expressed in this work are those of the authors and do not necessarily imply OTC policy or future service offerings.

7. REFERENCES

1. ANIDO, G.J. (1989), Traffic Control and Management Mechanisms for a Broadband Packet Network, *ITC Specialist Seminar*, Adelaide.
2. AUMANN, G. (1989), Source Policing in the Broadband-ISDN, *4th ATERB FPS Workshop*, Sydney.
3. CCITT Study Group XVIII Draft Recommendations (Jan. 1990), Geneva.
4. HEFFES, H. (1980), A Class of Data Traffic Processes - Covariance Function Characterization and Related Queuing Results, *Bell System Technical Journal*, Vol.59, No.6.
5. HEFFES, H., LUCANTONI, D.M. (1986), A Markov Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance, *IEEE Journal on Selected Areas in Communications*, Vol.4, No.6.
6. JOOS, P. and VERBIEST, W. (1989), A Statistical Bandwidth Allocation and Usage Monitoring Algorithm for ATM Networks, *ICC/89*, Boston.
7. KAWASHIMA, K. and SAITO, H. (1989) Teletraffic issues in ATM networks, *ITC Specialist Seminar*, Adelaide.
8. NEUTS, M.F. (1979), A versatile Markovian point process, *J. Appl. Prob.*, Vol.16, pp.764-779.
9. RAMASWAMI, V. (1980), The N/G/1 Queue and its Detailed Analysis, *Adv. Appl. Prob.*, Vol.12, pp.221-261.
10. ROSSITER, M.H. (1987), A switched Poisson model for data traffic, *Aust. Telecomm. Res.*, Vol.21, pp.53-57.
11. ROSSITER, M.H. (1988), The Switched Poisson Process and the SPP/G/1 Queue, *Aust. Telecomm. Res.*, Vol.22, pp.63-67.
12. VERBIEST, W., PINOO. L., VOETEN, B. (1988), Statistical Multiplexing of Variable Bit Rate Video Sources in Asynchronous Transfer Mode Networks, *Proceedings of Globecom'88*, pp. 7.2.1-7.2.6.
13. WALLMEIER, E. (1990) A Connection Acceptance Algorithm for ATM Networks Based on Mean and Peak Bitrates, submitted to *Int. J. Digital & Analogue Cabled Systems.*
14. WOODRUFF, G., ROGERS, R., RICHARDS, P. (1988), A Congestion Control Framework for High-speed Integrated Packetized Transport, *Proceedings of Globecom'88*, pp. 7.1.1-7.1.5.

15. WOODRUFF, G., KOSITPAIBOON, R., FITZPATRICK, G., RICHARDS, P. (1989), Control of ATM Statistical Multiplexing Performance, *ITC Specialist Seminar*, Adelaide.

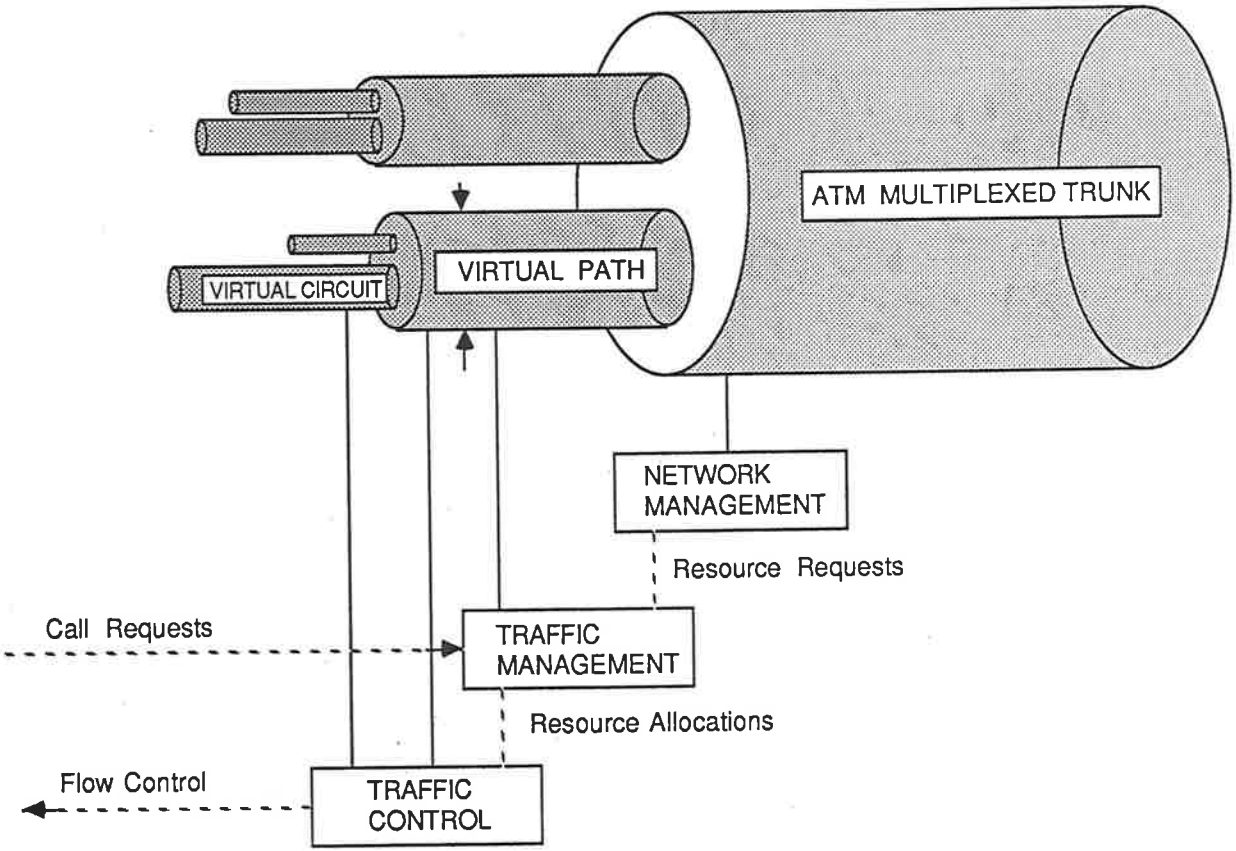


Figure 1

Network Management Architecture

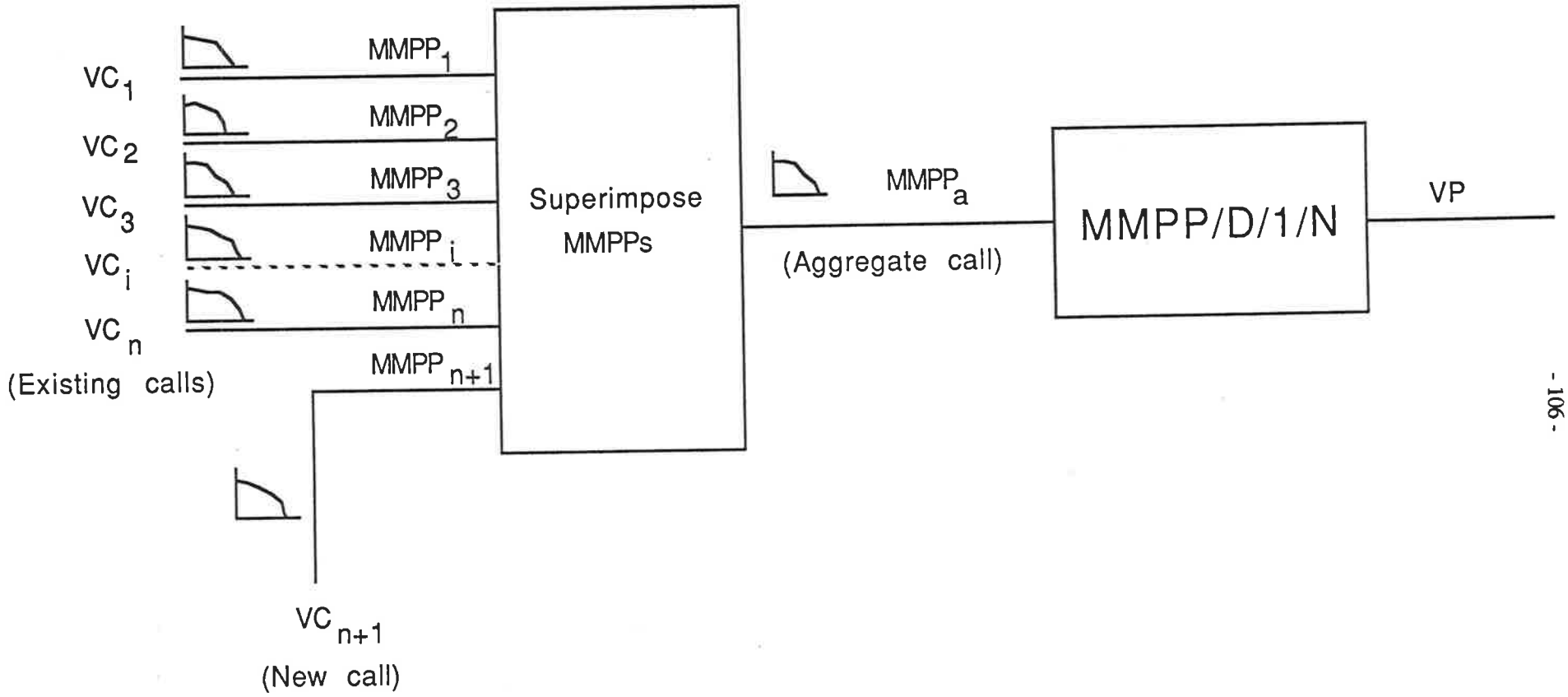


Figure 2

Call Admission Algorithm Scenario

Appendix B:

Connection Admission Control in ATM Networks

Paper presented at the Fifth Australian Teletraffic Research Seminar in Melbourne, 3-4 December 1990.

Connection Admission Control in ATM Networks

E. Dutkiewicz and G. Anido

Development Unit
OTC Limited
255 Elizabeth St
Sydney NSW 2000
New South Wales, AUSTRALIA

Phone: (02) 2873121
Fax: (02) 2873299
Acsnet: eric@otc.otca.oz.au

Abstract

A three level management architecture forms a basis for managing and controlling network resources in ATM networks. It utilises the concepts of virtual channels, virtual paths, and virtual networks. Call admission model using a 2-state Markov modulated Poisson process approximation for traffic streams is studied in order to obtain simple approximation algorithms.

1. INTRODUCTION

Broadband networks will be required to support a variety of services displaying a wide range of characteristics and demanding from the network different levels of performance. Of particular interest, at least in the early stages of the existence of broadband networks, will be the problem of supporting existing services which often require deterministic channels with stringent temporal and spatial requirements. Provision will also have to be made for other future services which may place different kinds of demand on the network. In order to deal with these various requirements traffic control and traffic management capabilities will have to be provided in such networks.

Admission control will form an important aspect of the control part of future broadband networks. A number of models have been proposed in the literature (see, for example [4]) in which source models are developed and studied in this context. In this paper a three level management architecture is presented for controlling and managing traffic in broadband networks. The problem of admission control is then studied under this framework. A queuing model for the connection admission control is developed and studied in order to obtain performance measures which can be used to describe the quality of service for connections.

In Section 2 of the paper we present a three level management architecture which allows a wide range of services to be accommodated and by means of which efficient use of the available resources can be made. In particular, use is made of the concepts of virtual channels and virtual paths which provide communication channels in ATM networks. The concept of virtual networks is defined wherein resource management of virtual paths is used as a means for providing guaranteed qualities of service in the network.

Using the above framework the problem of connection admission control is studied in Section 3. The basic function of the connection admission control is to ensure that connections are accepted provided there is enough resources to support them with guaranteed quality without degrading the quality of existing connections.

Results of studies based on the model presented in Section 3 are discussed in Section 4. A simple admission algorithm is proposed in Section 5.

2. MANAGEMENT ARCHITECTURE

The motivation for developing a management architecture for ATM networks is the need for a flexible framework under which network resources will be utilised in an efficient manner and which will guarantee to users required performance levels.

The management architecture which was first proposed in [1] and discussed further in [5] consists of three levels: ATM Resource Management, ATM Traffic Management, and ATM Traffic Control.

At the ATM Resource Management level virtual networks are set up and managed according to network demand. A virtual network is a group of virtual paths set up in the network to provide a specified quality of service. In particular, at this level use of resources in virtual networks is monitored and allocated resources are adjusted as demand for them changes.

At the ATM Traffic Management level resources allocated to individual virtual paths within virtual networks are managed. Connection admission control also take place at this level. This is based on the quality of service demanded by a connection, expected volume of traffic in the connection, and the current state of the virtual path into which the connection is to be admitted. Grade of service (in terms of call blocking probability) may be also taken into account in the admission process.

At the ATM Traffic Control level actions are taken to ensure that traffic volumes in individual virtual channels and virtual paths do not exceed allowed thresholds. This can be carried out by source policing traffic entering virtual channels and virtual paths. In virtual paths level traffic control is required in order to provide separation between different virtual paths and in this manner to guarantee quality of services provided by those virtual paths.

The management architecture showing the above levels is depicted in Figure 1.

3. CONNECTION ADMISSION CONTROL

Quality of service and grade of service are two criteria on which connection admission into ATM networks may be based. In order to distinguish these two terms it may be convenient to associate quality of service with cell level performance and grade of service with call level performance. As a result quality of service is often specified in terms of cell loss probability and cell delay and grade of service in terms of call blocking probability.

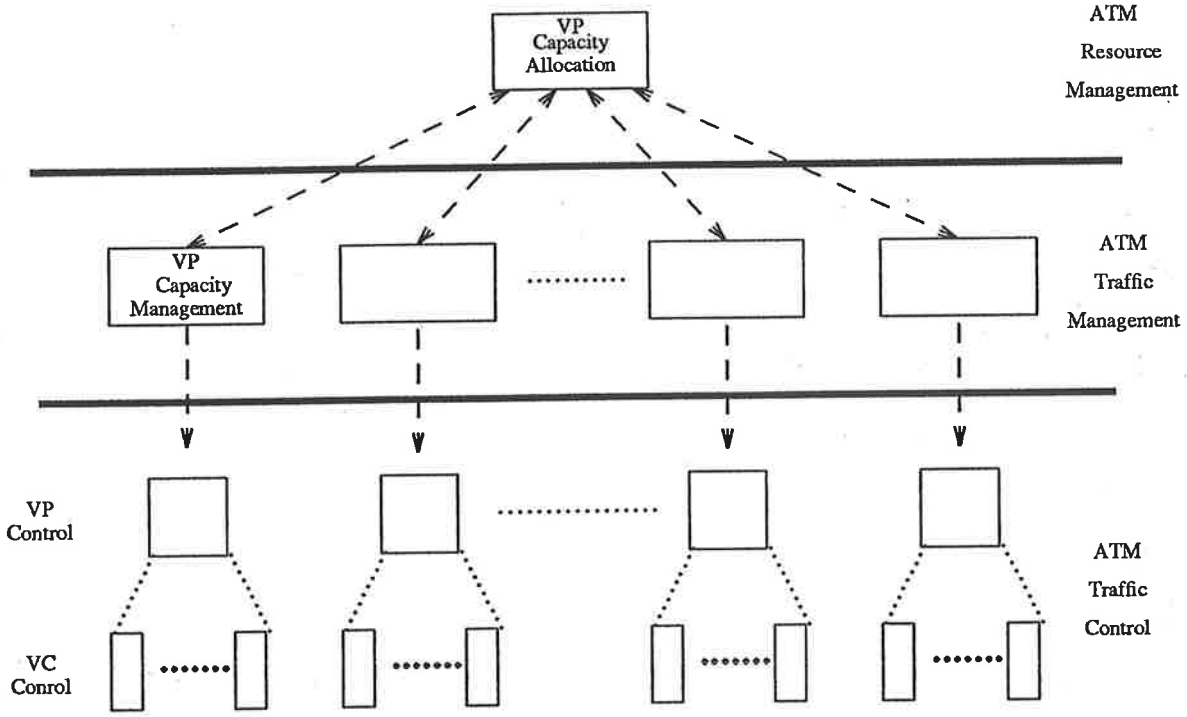


Figure 1: Traffic Management and Control Architecture for ATM

Approaches taken in the literature concentrate on the quality of service aspects of the connection admission control. A common approach relies on the knowledge of connection interarrival time distributions. The statistical parameters used to describe such distributions include average bit rate, bit rate variance, and peak bit rate. In [8] average bit rate and bit rate variance of individual connections is used in the admission algorithm. It is assumed that when a sufficient number of such connections is multiplexed together the resulting bit rate distribution will approach a normal distribution. This will then allow the determination of the maximum allowable load which can be permitted for a given cell loss performance.

A similar approach is taken in [16]. In this case the acceptance decision is based on the mean bit rates and peak bit reates of the new and existing connections without the need for the bit rate variance.

The approaches based on the knowledge of bit rate distributions do not take into account the fluctuations of bit rates in time. Thus a model which incorporates a measure of burstiness might be more useful.

3.1 Connection Admission Model

In this model the criterion for connection admission is the quality of service requested by the connection. The grade of service requested by all connections is assumed identical.

Quality of service measure is measured in terms of cell loss probability. Cell delay performance, which depends on the size of the network and the route through the network, is regarded to be of less importance. This is particularly the case given that ATM buffers are expected to be small [8].

Using the idea of virtual networks, each virtual path belonging to the same virtual network will have to support the same quality of service. Connections which require different qualities of service are assigned into virtual paths within distinct virtual networks corresponding to the required qualities of service. In order to ensure that distinct qualities of service are provided by virtual network, traffic streams belonging to different virtual networks will have to separated. One method which can be used to carry out such separation is through peak limiting the capacity assigned to virtual paths within different virtual networks. Capacity utilised by each virtual path will then be policed using a policing mechanism placed at the origin of each virtual path. As a result, the quality of service provided by a virtual network will be determined by the policing mechanisms associated with all virtual paths belonging to the virtual network and by the connection admission mechanism.

3.2 Queuing Model for Connection Admission Control

The policing mechanism at the origin of a virtual path will ensure that the peak bandwidth of the traffic carried in the virtual path does not exceed the allowed threshold. One method of implementing such a policing mechanism can be achieved by means of a leaky bucket [2].

A leaky bucket for policing virtual paths can be modelled as a short queue with a deterministic server. The service rate of the queue sets the allowed peak bit rate of the traffic carried by the virtual path and the length of the queue should be set so as to ensure separation between traffic carried by different virtual paths. However, setting the length of the queue to $N = 1$ so as to achieve complete separation would not allow short bursts of cells on the transmission line (due to jitter introduced by multiplexing equipment) to be accommodated. As a result the exact setting of the queue length will have to be a compromise between the above requirements.

The queue modelling the leaky bucket as the policing mechanism of the virtual path can be represented as G/D/1/N queue using Kendall notation where G represents a general input process, D a deterministic server, and N the length of the queue including the server.

The choice of the arrival process into the virtual path representing individual connections has been influenced by need for a process which can be used to model a wide range of different traffic streams and at the same time one which will allow tractable analysis to be carried out when applied to a queuing system. An arrival process which has attracted a lot of attention on account of satisfying the above criteria is the 2-state Markov modulated Poisson process.

A 2-state Markov modulated Poisson process (2-state MMPP) is a special case of the more general Markov modulated Poisson process (MMPP). An MMPP is a Poisson process whose instantaneous rate is a stationary random process which evolves according to an irreducible m state Markov chain. It can be characterised by an infinitesimal generator for the underlying Markov process **R**

and an arrival matrix Λ given by:

$$\mathbf{R} = \begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & r_{1,4} & \cdot & \cdot & r_{1,m} \\ r_{2,1} & r_{2,2} & r_{2,3} & r_{2,4} & \cdot & \cdot & r_{2,m} \\ r_{3,1} & r_{3,2} & r_{3,3} & r_{3,4} & \cdot & \cdot & r_{3,m} \\ r_{4,1} & r_{4,2} & r_{4,3} & r_{4,4} & \cdot & \cdot & r_{4,m} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \\ r_{m,1} & r_{m,2} & r_{m,3} & r_{m,4} & \cdot & \cdot & r_{m,m} \end{bmatrix}$$

$$\Lambda = diag\,(\lambda_1,..,\lambda_m)$$

In the case of the 2-state MMPP the above reduce to:

$$\mathbf{R} = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

The parameters λ_1, λ_2 required to specify Λ correspond to the intensity of the Poisson processes in states 1 and 2 of the underlying Markov process respectively, whereas r_1, and r_2 correspond to the transition rates of the Markov process.

The solution of the 2-state MMPP/D/1/K can be obtained by following analysis for more general processes as presented in the literature such as for the N/G/1 queue found in [13] and equivalent BMAP/G/1 queue found in [10]. An outline of an algorithm which has been implemented and used to solve the 2-state MMPP/D/1/K queue has been given in the Appendix.

Solution of the above queuing system allows modelling of the policing mechanism given an aggregate traffic stream consisting of all traffic streams entering a given virtual path. Connection admission will, however, have to deal with individual connections and as a result a method of obtaining a superposition of individual connections which can then be applied to the queue is needed. Two such superposition methods were investigated. In the first method for obtaining a superposition of 2-state MMPPs developed in [7] (Method I) the following characterising parameters have been used:

m - mean arrival rate,

v - variance of arrival rate,

μ_3 - third moment of arrival rate,

τ - time constant of the arrival rate.

When n 2-state MMPPs are superimposed together characterising parameters of the superimposed stream are given by:

$$m = \sum_{i=1}^{n} m_i$$

$$v = \sum_{i=1}^{n} v_i$$

$$\mu_3^* = \sum_{i=1}^{n} \mu_{3i}^*$$

$$\tau = \sum_{i=1}^{n} \frac{v_i}{v} \tau_i$$

where

$$\mu_3^* = \mu_3 - 3mv - m^3$$

The defining parameters for the 2-state MMPP representing the superimposed stream can be obtained by inversion of the above equations as given in [7].

In the second method proposed in [14] (Method II) the arrival process is characterised by the following parameters:

λ - mean arrival rate,

c^2 - squared coefficient of variation of interarrival times,

$Z(\infty)$ - asymptotic variance to mean ratio of the number of arrivals,

$C(\infty)$ - asymptotic covariance of the number of arrivals.

When n 2-state MMPP streams are superimposed together the set of the characterising parameters for the superimposed stream is given by:

$$\lambda = \lambda_1 + ... + \lambda_n$$

$$Z(\infty) = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} Z_i(\infty)$$

$$C(\infty) = \sum_{i=1}^{n} C_i(\infty)$$

$$c^2 = E(T)2\lambda - 1$$

where E(T) is the mean of the forward recurrence time and is given by:

$$E(T) = \frac{1 + c^2}{2\lambda}$$

The set of the parameters defining the 2-state MMPP can be obtained from the characterising parameters after some algebraic manipulation as given in [14].

In order to determine their suitability for the connection admission algorithm the two superposition methods were applied to the problem of superposition of n voice streams with the resulting aggregate stream used as an input to the queuing system described above. A voice stream was modelled by an interrupted Poisson process (IPP) based on the results presented in [15] and [14]. The following parameters were used to define each stream: $\lambda_1 = 54.44$ *cells/sec*, $\lambda_2 = 0$, $r_1 = 2.268$ sec^{-1}, and $r_2 = 1.532$ sec^{-1}.

Figures 2 and 3 show the defining parameters of the superimposed stream obtained by using the two methods. In both figures Method II used stream superposition in groups of 10 streams. The effect of stream grouping used in Method II is particularly evident in Figure 3. Method I is insensitive to such groupings. This means that superimposing an additional stream to an existing aggregate will produce the same result as the superposition of all individual streams. This becomes an important issue when the implementation of the algorithm is considered. In Method II the computationally expensive step is associated with the calculation of the coefficient of variation c^2 where the cost increases as 2^n with the number of different streams n.

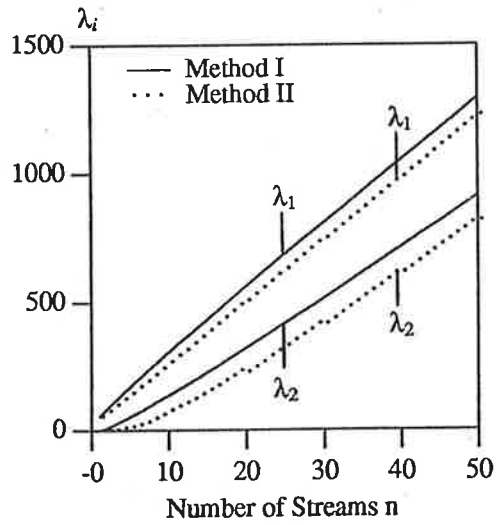


Figure 2: Component Rates λ_i for Superposition

The effect of grouping streams in groups of 10 and applying Method II to carry out the superposition as opposed to using Method II

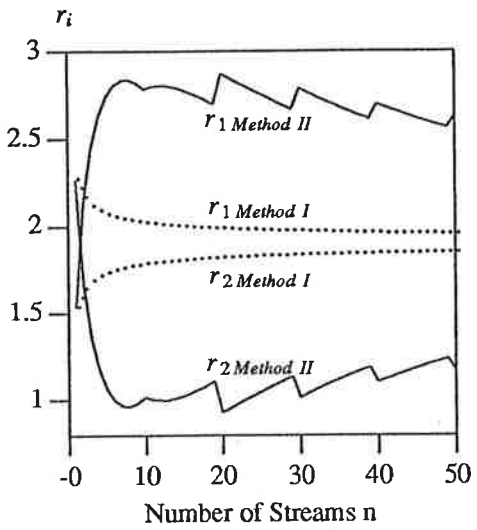


Figure 3: Phase Rates r_i for Superposition

without grouping is shown in Figure 4 for the resultant transition rates r_1 and r_2 of the superimposed stream.

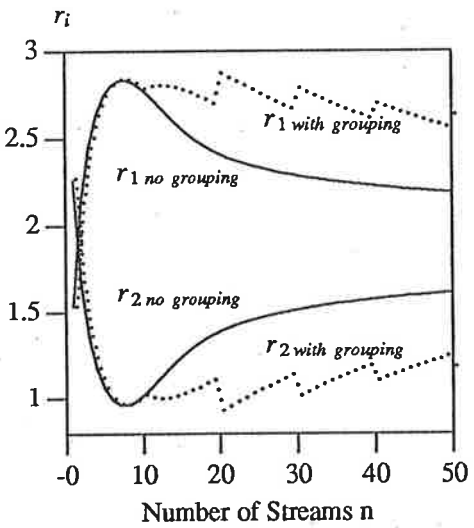
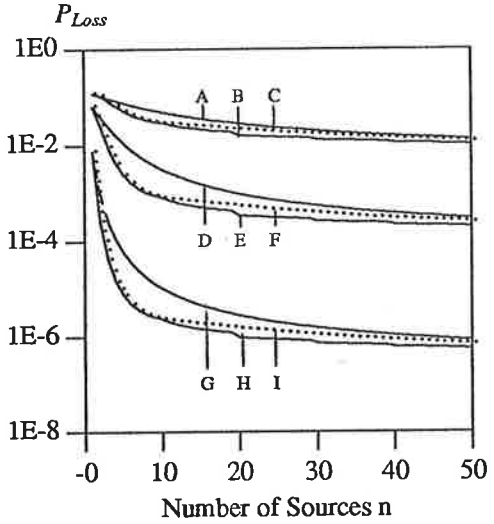


Figure 4: Effect of Grouping on Phase Rates in Method II

The two superposition methods were also used to produce a superimposed stream which was subsequently applied to the 2-state MMPP/D/1/K queue. Figures 5 and 6 show the loss probability produced for two different queue length K and a number of different utilisation levels. Separate curves were obtained for Method II with stream grouping in groups of 10 and without stream grouping. The resultant curves indicate that stream grouping in Method II underestimates the resultant loss probability as compared to Method II used without stream grouping. Note also that

Method I gives the most conservative results, however, as the number of streams increases the results obtained using the two methods converge. Queueing performance criteria obtained for the two superposition methods must be treated with caution. They show relative performance as there is no baseline against which they could both be judged. It should also be noted that traffic characterisation which will allow for accurate queueing performance prediction is a topic of some debate. As a result, the choice of the superposition method for subsequent studies was based on the ease of implementation. In this respect Method I was found more efficient.



Legend:

A : Method I, ρ =0 .8

B : Method II with grouping, $\rho = 0.8$

C : Method II without grouping, $\rho = 0.8$

D : Method I, $\rho = 0.6$

E : Method II with grouping, $\rho = 0.6$

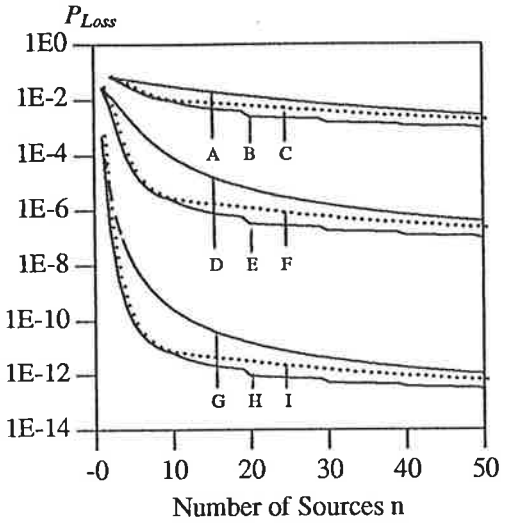
F : Method II without grouping, $\rho = 0.6$

G : Method I, $\rho = 0.4$

H : Method II with grouping, $\rho = 0.4$

I : Method II without grouping, $\rho = 0.4$

Figure 5: Loss probability for K=10 using Methods I and II.



Legend as for Figure 5.

Figure 6: Loss probability for K=20 using Methods I and II.

4. SUPERPOSITION OF HETEROGENEOUS STREAMS

A virtual path which is associated with quality of service rather than being service specific can be expected to carry a number of dissimilar traffic streams. Algorithms for connection admission control will have to accommodate such wide range of services. The algorithm for solution of the 2-state MMPP/D/1/K queue was used to investigate performance of such queues when two types of different traffic streams were used. The two traffic types representing voice and video services have been chosen. The voice stream model has been presented earlier. The video stream was modelled as an IPP based on the studies in [11] with the resultant 2-state MMPP parameters as follows: $\lambda_1 = 11.1 * 10^3$ *cells/sec*, $\lambda_2 = 0$, $r_1 = 3.25$ sec^{-1}, and $r_2 = 0.64$ sec^{-1}.

The two stream types were applied to the 2-state MMPP/D/1/K queue and boundaries for the resultant loss probability of 10^{-8} were obtained for different queue sizes K as shown in Figure 7. Also, having fixed the queue size to K=10, the boundaries for the same loss probability have been obtained for different queue service rates (which correspond to different output link capacities). These are shown in Figure 8.

The acceptance boundaries shown in Figures 7 and 8 display near linearity, the consequence of which could be exploited in producing approximate acceptance algorithms. Simple

approximations based just on the end points of the acceptance boundaries may not be possible as the boundaries display concavity. Further studies will be carried out to determine the amount of concavity for different configurations. Similar results concerning the linearity of acceptance boundaries have been noted in [6] where a bufferless network was assumed with traffic modelled as interrupted deterministic streams and the statistical behaviour was studied using a large deviation approximation.

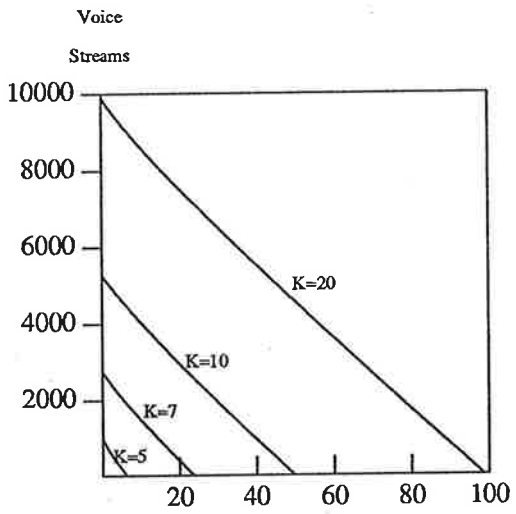


Figure 7: Acceptance Regions for Different Queue Sizes K

The noted linearity of the acceptance boundaries could be used in developing simple connection admission algorithms utilising the concept of effective bandwidth. (For a survey of such methods see [12]).

Another important consideration in developing connection admission algorithms is the sensitivity of bandwidth requirement for a connection as the quality of service changes. In order to study this problem voice streams as defined previously were applied to a deterministic queue. The effective bandwidth defined as the available capacity over the number of streams carried, was obtained for such systems against loss probability. Figure 9 shows the resultant curves obtained for different values of queue length K.

Curves obtained in Figure 9 suggest that as the queue length increases changes in many orders of magnitude to cell loss result in very small changes to the effective bandwidth requirement. On the other hand, it should be noted that only a small

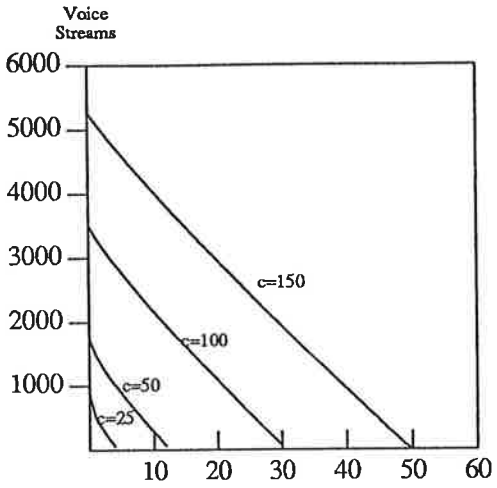


Figure 8: Acceptance Regions for Different Queue Service Rates c

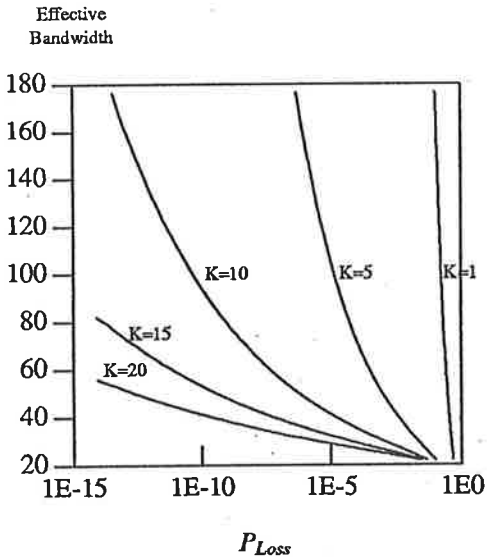


Figure 9: Effect of Effective Bandwidth on Loss Probability

change in effective bandwidth may result in changes to cell loss probability of many magnitudes. This could have important consequences on the sensitivity of the quality of service provided in a virtual path.

5. CONNECTION ADMISSION CONTROL ALGORITHM

A simple connection admission algorithm based on the models presented earlier has been developed. Connections requesting a particular quality of service can be admitted into a virtual

path according to the algorithm steps depicted in Figure 10. Call parameters are used to determine the set of characterising parameters as well as connection type. It is assumed that there is a finite number of connection types possible and look-up tables can be used to match connection parameters to the required connection type if the connection type is not specified explicitly. In the next step in the algorithm the new connection is superimposed with the existing aggregate of connections carried by the virtual path. The resultant stream is then applied to the queuing system which models the virtual path and the queuing system is solved for cell loss probability. If the resultant cell loss probability exceeds that allowed for the virtual path the new connection is rejected, otherwise the new connection is accepted and the characterising parameters of the new aggregate connection in the virtual path are stored.

The process of removing connections from the virtual path is depicted in Figure 10. It involves only the calculation of the new aggregate stream without the need for solving the queuing system. The first step in the algorithm involves determination of the characterising parameters of the connection to be removed. This information can be obtained from a look-up table if the type of the connection is known. Next the new aggregate for the virtual path with the required connection removed is obtained and the characterising parameters for the new aggregate are saved.

Removal of an individual stream from a superimposed stream can be carried out in a simple fashion if Method I for stream superposition is used. Let m, ν, μ_3, τ and $m_i, \nu_i, \mu_{3_i}, \tau_i$ denote the characterising parameters of the existing aggregate and the connection to be removed respectively. Then the characterising parameters of the new aggregate with the connection removed can be obtained from:

$$m_n = m - m_i,$$

$$\nu_n = \nu - \nu_i,$$

$$\mu_{3n}{}^* = \mu_3{}^* - \mu_{3i}{}^*,$$

$$\tau_n = \frac{1}{\nu - \nu_i}(\nu\ \tau - \nu_i\ \tau_i).$$

The main purpose of the above algorithm is to provide a basis for development and comparison of simple connection admission algorithms such as those using the idea of the effective bandwidth. It should also be noted that the above steps can form the basis of an algorithm which could be

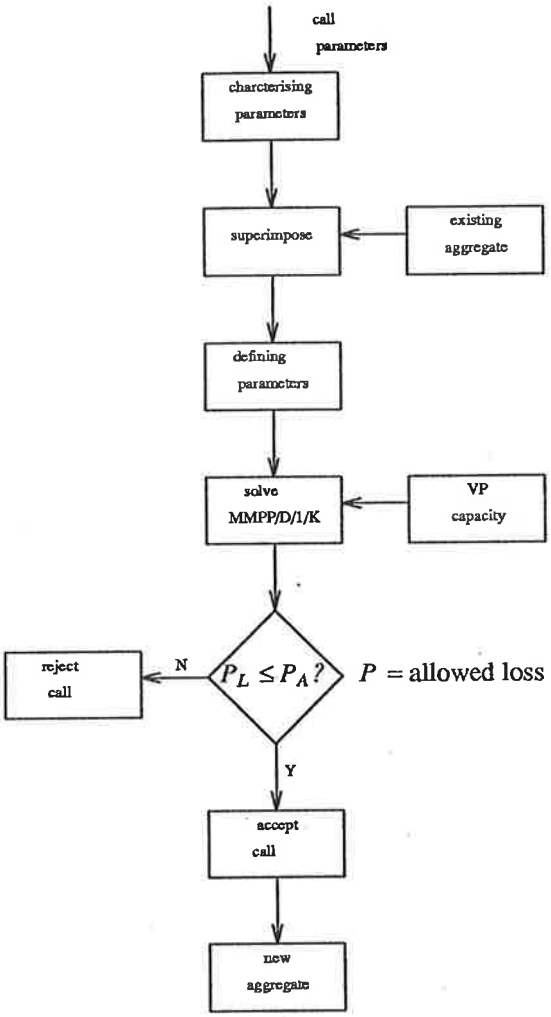


Figure 10: Algorithm Steps for Adding New Connections

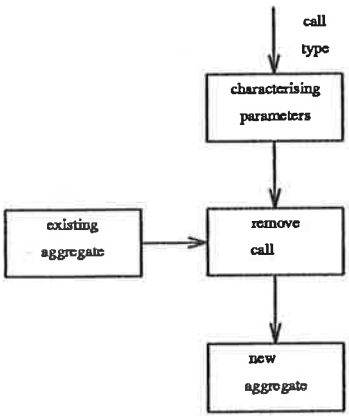


Figure 11: Algorithm Steps for Removing Old Connections

used in practice. However, the step involving the solution of the queueing system may impair the performance of the algorithm and as a result

simplifications of this procedure may be required.

6. CONCLUSION AND FURTHER WORK

A versatile three level management architecture for controlling and managing traffic in future broadband networks has been proposed.

A queuing model for connection admission control has been presented and a connection admission control algorithm based on the model has been proposed.

The main purpose of the above algorithm is to act as a vehicle for further studies of simple connection admission algorithms. In particular, studies of acceptance regions with a wide variety of different traffic types for development of algorithms which use the concept of the effective bandwidth will be carried out.

7. APPENDIX

A matrix algorithmic solution of the 2-state MMPP/D/1/K queue from which queue length distribution and loss probability can be obtained can proceed according to the following steps. Let $\tilde{\mathbf{P}}(x)$ denote the transition probability matrix for a queue with an infinite number of spaces. This matrix can be expressed as

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

where $\tilde{\mathbf{A}}_n(x)$ and $\tilde{\mathbf{B}}_n(x)$ are 2×2 matrices. The stationary transition probability matrix $\mathbf{P} = \tilde{\mathbf{P}}(\infty)$ can be used to obtain the stationary queue-length density $\mathbf{x}$ at departure epochs. The defining system of equations is:

$$\mathbf{x}\,\tilde{\mathbf{P}}(\infty) = \mathbf{x}, \quad \mathbf{x}\,\mathbf{e} = 1.$$

The effect of a finite queue is the truncation of the state space resulting in the stationary transition probability matrix of the Markov chain which becomes:

$$\tilde{\mathbf{P}} = \begin{bmatrix} \tilde{\mathbf{B}}_0 & \tilde{\mathbf{B}}_1 & \tilde{\mathbf{B}}_2 & . . & \tilde{\mathbf{B}}_{N-2} & \sum_{k=N-1}^{\infty} \tilde{\mathbf{B}}_k \\ \tilde{\mathbf{A}}_0 & \tilde{\mathbf{A}}_1 & \tilde{\mathbf{A}}_2 & . . & \tilde{\mathbf{A}}_{N-2} & \sum_{k=N-1}^{\infty} \tilde{\mathbf{A}}_k \\ 0 & \tilde{\mathbf{A}}_0 & \tilde{\mathbf{A}}_1 & . . & \tilde{\mathbf{A}}_{N-3} & \sum_{k=N-2}^{\infty} \tilde{\mathbf{A}}_k \\ . & . & . & . . & . & . \\ . & . & . & . . & . & . \\ 0 & 0 & 0 & . . & \tilde{\mathbf{A}}_0 & \sum_{k=1}^{\infty} \mathbf{A}_k \end{bmatrix}$$

An efficient algorithm for the computation of matrices $\mathbf{A}_n$ has been presented in [9]. Computation of the stationary probability vector $\mathbf{x}$ at departure epochs can then be carried out following simple matrix analysis [3]. Queue length distribution $\mathbf{y}$ at an arbitrary time t can be obtained next by application of the key renewal theorem [13]. The blocking probability for the queue can be then obtained as:

$$P_{block} = 1 - \sum_{n=0}^{N-1} \mathbf{y}_n \mathbf{e}.$$

8. ACKNOWLEDGEMENT

The permission of the Managing Director of OTC to publish this work is gratefully acknowledged. The views and opinions expressed in this work are those of the authors and do not necessarily imply OTC policy or future service offerings.

9. REFERENCES

1. Anido G., "Traffic Control and Management Mechanism for a Broadband Packet Network", ITC 13, Adelaide, 1989.
2. Aumann G., "Source Policing in the Broadband-ISDN", 4th ATERB FPS Workshop, Sydney, July 1989.
3. Blondia C., "The N/G/1 Finite Capacity Queue", Commun. Statist.-Stochastic Models, 5(2), 273-294 (1989).
4. Dittmann L., and Jacobsen S.B., "Statistical Multiplexing of Identical Bursty Sources in an ATM Network", IEEE GLOBECOM'88.
5. Dutkiewicz E and Anido G.J, "Traffic Management and Control in Broadband Networks", 5th ATERB FPS Workshop, Melbourne, July 1990.

6. Griffiths T.R., "Analysis of Connection Acceptance in Asynchronous Transfer Mode Networks", British Telecom Research Laboratories, 1990.

7. Heffes H., "A Class of Data Traffic Processes - Covariance Function Characterization and Related Queueing Results", BSTJ, vol. 59, no. 6, July-August 1979.

8. Joos P., Verbiest W., "A Statistical Bandwidth Allocation and Usage Monitoring Algorithm for ATM Networks", ICC/89, Boston, June 1989.

9. Lucantoni D.M., "Efficient Algorithms for Solving the Non-linear Matrix Equations Arising in Phase Type Queues", Commun. Statist-Stochastic Models, 1(1), 29-51 (1985).

10. Lucantoni D.M., "New Results on the Single Server Queue with a Batch Markovian Arrival Process", AT&T Bell Laboratories, 1989.

11. Malgris B., et al, "Performance Models of Statistical Multiplexing in Packet Video Communications", IEEE Transactions on Communications, vol. 36, no. 7, July 1988.

12. Mase K., and Shioda S., "Real-Time Network Management for ATM Networks", to be presented at the 13th ITC, Copenhagen 1991.

13. Ramaswami V., "The N/G/1 Queue and its Detailed Analysis", Adv. Appl. Prob., vol 12, pp. 222-261, Mar 80.

14. Rossiter M.H., "Sojourn Time Theory and the Switched Poisson Process", Telecom Australia Research Laboratories Report 7835, 1986.

15. Sriram K., Whitt W., "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data", IEEE Journal on Selected Areas in Communications, vol. SAC-4, no. 6, September 1986.

16. Wallmeier E., "A Connection Acceptance Algorithm for ATM Network Based on Mean and Peak Bitrates", submitted to Int. J. Digital and Analogue Cabled Systems.

Appendix C:

Queue Length Distribution in Continuous Time

The queue length distribution at an arbitrary time required in the queueing analysis of Chapter 4 can be obtained in the following manner. Define the queue length distribution at an arbitrary time by:

$$y(i,j) = \lim_{t\to\infty} P\left\{X(t)=i, J(t)=j \mid X(0)=i', J(0)=j'\right\} \quad i \geq 0, 1 \leq j \leq m, \; t \geq 0. \qquad \text{(C.1)}$$

Probability $K_{(i',j')(i,j)}(t) = P\left\{X(t)=i, J(t)=j \mid X(0)=i', J(0)=j'\right\}$ can be obtained as the solution of the Markov renewal equation [10] applied to the queue process. The Markov renewal equation can be stated in this case as:

$$K_{(i',j')(i,j)}(t) = G_{(i',j')(i,j)}(t) + \sum_{k=1}^{m} \sum_{v=0}^{i} \int_0^t K_{(v,k)(i,j)}(t-x)\, d\vec{P}_{(i',j')(v,k)}(x) \qquad \text{(C.2)}$$

where $\vec{P}_{(i',j')(i,j)}(x)$ is the semi-Markov matrix for the queue length process (cf Equation 4.11) and

$$G_{(i',j')(i,j)}(t) = P\left\{X(t)=i, J(t)=j, t < x \mid X(0)=i', J(0)=j'\right\} \qquad \text{(C.3)}$$

and x is the time of last renewal since $t=0$ (corresponding to the last departure epoch since $t=0$). The first part of Equation C.2 covers the case when the first departure from the queue since $t=0$ occurs at time x, and the second part of Equation C.2 takes into account all the cases in which there is one or more

departures from the queue since $t = 0$ and before time x.

The solution to this equation is then given by [10]:

$$K_{(i',j')(i,j)}(t) = \sum_{k=1}^{m} \sum_{v=0}^{i} \int_0^t G_{(v,k)(i,j)}(t-x)\, dR_{(i',j')(v,k)}(x) \qquad \text{(C.4)}$$

where $R_{(i',j')(i,j)}(t) = E[M_{(i,j)}(t) \mid X_0 = (i',j')]$ is the expected number of times that the process enters state (i,j) in the interval [0,t] given that at $t = 0$ it was in state (i',j'). $M_{(i,j)}(t)$ is the number of times the process enters state (i,j) in [0,t].

This leads to the following expressions for $K_{(i',j')(i,j)}(t)$:

$$K_{(i',j')(0,j)}(t) = \sum_{k=1}^{m} \int_0^t P_{kj}(0,t-x)\, dR_{(i',j')(0,k)}(x) \qquad \textit{for } i = 0 \qquad \text{(C.5a)}$$

and

$$K_{(i',j')(i,j)}(t) = \sum_{k=1}^{m} \int_0^t \int_0^{t-x} (1-\tilde{H}(t-x-u)) \sum_{v=1}^{i} \sum_{p=1}^{m} [d\tilde{U}_v(u)]_{kp} P_{pj}(i-v, t-x-u)\, dR_{(i',j')(0,k)}(x)$$

$$+ \sum_{k=1}^{m} \sum_{v=1}^{i} \int_0^t (1-\tilde{H}(t-x)) P_{kj}(i-v, t-x)\, dR_{(i',j')(v,k)}(x) \quad \textit{for } i \geq 1 \qquad \text{(C.5b)}$$

where $P_{jk}(n,t)$ is given by: (cf. Equation 4.2 in Chapter 4)

$$P_{j'j}(n,t) = P\left\{ N(t) = n, J(t) = j \mid N(0) = 0, J(0) = j' \right\}$$

where $N(t)$ indicates the number of arrivals in time t and $J(t)$ the phase of the arrival process at time t.

Also $\tilde{\mathbf{U}}_v(x) = [\int_0^x \mathbf{P}(0,y)dy]\, \mathbf{D}_v$. The (j,k)th entry of $\tilde{\mathbf{U}}_v(x)$ is the conditional probability, given $J(0) = j$, that the first arrival to an empty queue occurs at or before x and is of group size v, and the phase of the process at the time of the first arrival is k.

Equation C.5a corresponds to the case where the queue is empty at time t. The last departure from the queue occurring at time x leaves the queue empty and there are no arrivals in time $t-x$. Equation C.5b corresponds to the case where the queue is not empty at time t. The first term of Equation C.5b refers to the case where the departure at time x leaves the queue empty and the first arrival of group size v to the empty queue occurring at time $x+u$ is followed by $i-v$ arrivals in time $t-x-u$. The second term of Equation C.5b refers to the case where there are v customers in the queue at the last departure epoch x and additional $i-v$ arrivals arrive in time $t-x$ before the next departure.

The limiting behaviour of $K_{(i',j')(i,j)}(t)$ can be obtained by application of the key renewal theorem. The key renewal theorem, which is a generalisation of the renewal theorem [21], states [10]:

$$\lim_{t\to\infty}\int_0^t G_{(v,k)(i,j)}(t-x)dR_{(i',j')(v,k)}(x) = m^{-1}(v,k)\int_0^\infty G_{(v,k)(i,j)}(x)dx \qquad \text{(C.6)}$$

where $m^{-1}(v,k)$ is the mean recurrence time of (v,k) and $R_{(i',j')(v,k)}(t)$ and $G_{(v,k)(i,j)}(t)$ are as defined for Equation C.4 and Equation C.3 respectively.

Applying the key renewal theorem to each term in Equation C.5 gives:

$$y(0,j) = \sum_{k=1}^{m} m^{-1}(0,k)\int_0^\infty P_{kj}(0,t)dt \qquad \text{for } i=0 \qquad \text{(C.7a)}$$

and

$$y(i,j) = \sum_{k=1}^{m} m^{-1}(0,k)\int_0^\infty\int_0^t (1-\tilde{H}(t-x))\sum_{v=1}^{i}\sum_{p=1}^{m}[d\tilde{U}_v(x)]_{kp}P_{pj}(i-v,t-x)\,dt$$
$$+\sum_{k=1}^{m}\sum_{v=1}^{i} m^{-1}(v,k)\int_0^\infty (1-\tilde{H}(t))P_{kj}(i-v,t)dt \quad \text{for } i\geq 1 \qquad \text{(C.7b)}$$

The mean recurrence time of (v,k) is related to the stationary queue length distribution x_{vk} at departure epochs by [46]:

$$m^{-1}(v,k) = \lambda x_{vk} \tag{C.8}$$

Substituting the above in Equation C.7a, noting that $\int_0^\infty P_{kj}(0,t)dt = -D_0^{-1}(k,j)$, and using matrix notation gives:

$$\mathbf{y}_0 = -\lambda \mathbf{x}_0 \mathbf{D}_0^{-1} \tag{C.9a}$$

Putting Equation C.8 in Equation C.7b and expressing the result in matrix notation gives:

$$\mathbf{y}_i = -\lambda \mathbf{x}_0 \int_0^\infty \int_0^t (1 - \bar{H}(t-x)) \sum_{v=1}^{i} [d\tilde{\mathbf{U}}_v(x) \mathbf{P}(i-v, t-x)\, dt$$

$$+ \lambda \sum_{v=1}^{i} \mathbf{x}_v \int_0^\infty (1 - \bar{H}(t)) \mathbf{P}(i-v,t)dt \quad \textit{for } i \geq 1$$

This can be simplified further [46] to give:

$$\mathbf{y}_i = \sum_{v=1}^{i} \lambda [\mathbf{x}_0 \mathbf{U}_v(0) + \mathbf{x}_v] \int_0^\infty (1 - \bar{H}(t)) \mathbf{P}(i-v,t)dt \quad \textit{for } i \geq 1. \tag{C.9b}$$

where $\mathbf{U}_v(s)$ is the Laplace-Stieltjes transform of $\tilde{\mathbf{U}}_v(x)$ given by:

$$\mathbf{U}_v(s) = \int_0^\infty e^{-sx} d\tilde{\mathbf{U}}_v(x)$$

Furthermore, $\mathbf{U}_v(0) = -\mathbf{D}_0^{-1}\mathbf{D}_v$, and the (k,j)th entry of $\mathbf{U}_v(0)$ can be interpreted as the conditional probability that an idle period ends with the arrival which has a group size of v and the phase of the arrival process is j, given that at the beginning of the idle period the arrival phase was k.

Appendix D:

Method for Computation of A_n Matrices

The numerical computation of matrices A_n, which form the elements of the transition probability matrix of the semi-Markov process describing the BMAP/G/1/N queue, and which are therefore necessary for the computation of the stationary queue length distribution at departure epochs x, can be accomplished using the technique of uniformisation.

The basic idea of uniformisation [19] is as follows. Consider a continuous time Markov chain with Q as its infinitesimal generator. Then the Chapman-Kolmogorov equation is given by:

$$P(t)' = P(t)Q$$

Solutions to this equation are easy to obtain if $q_i = \sum_{j \neq i} q_{ij}$ (leaving rate of state i) is equal for all i. Now, these rates can be made equal by introducing into the system null-events. This can be done by choosing a value $q \geq \max q_i$ for the new leaving rates of all states. The states of the system following each jump form, as a result, a discrete-time Markov chain, where the rate of each jump is q and each jump is independent of the state. Transition probabilities p_{ij} of the chain are given by:

$$p_{ij} = \frac{q_{ij}}{q} \quad for \ i \neq j$$

$$p_{ii} = 1 - \frac{q_i}{q}$$

Then the probability of being in state j after k jumps can be obtained as:

$$\pi_j^k = \sum_{i=1}^{n} \pi_i^{k-1} p_{ij} \quad ,k > 0$$

and the probability of being in state j at time t can be then obtained as:

$$\pi_j(t) = \sum_{k=0}^{\infty} \pi_j^k \, p(k;qt) \tag{D.1}$$

where

$$p(k;qt) = P(k \text{ jumps until } t) = e^{-qt} \frac{(qt)^k}{k!}$$

Now, in the case of the BMAP process we have:

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{j=0}^{\infty} e^{\theta t} \frac{(\theta t)^j}{j!} \mathbf{K}^j$$

where $\mathbf{K} = \mathbf{I} + \theta^{-1}\mathbf{Q}$. This can be expressed further as [37]:

$$\mathbf{P}(n,t) = \sum_{j=0}^{\infty} e^{-\theta t} \frac{(\theta t)^j}{j!} \mathbf{K}_n^j \tag{D.2}$$

where

$$\theta \geq \max(-\mathbf{D}_0)_{ii}$$

$$\mathbf{K}_0^{j+1} = K_0^j (\mathbf{I} + \theta^{-1}\mathbf{D}_0)$$

$$\mathbf{K}_n^{j+1} = \theta^{-1} \sum_{i=0}^{n-1} \mathbf{K}_i^j \mathbf{D}_{n-i} + \mathbf{K}_n^j (\mathbf{I} + \theta^{-1}\mathbf{D}_0)$$

with $\mathbf{K}_0^0 = \mathbf{I}, \mathbf{K}_n^0 = 0, n \geq 1$

In this case the discrete-time Markov chain is formed by introducing a (uniformising) Poisson process with rate θ. Matrix $\mathbf{K}_n$ is the transition probability matrix of this chain. The diagonal elements of this matrix can be interpreted as probabilities corresponding to dummy jumps from a state to itself.

Now let

$$\gamma_n = \int_0^\infty e^{-\theta t} \frac{(\theta t)^n}{n!} d\tilde{H}(t) \tag{D.3}$$

be the probability of n arrivals of the uniformising Poisson process in one service time.

Also recall (Equation 4.12 in Chapter 4) that matrices $\mathbf{A}_n$ can be obtained as:

$$\mathbf{A}_n = \int_0^\infty \mathbf{P}(n,t) d\tilde{H}(t)$$

Substituting for $\mathbf{P}(n,t)$ (Equation D.2) gives:

$$\mathbf{A}_n = \sum_{j=0}^{\infty} \int_0^\infty e^{-\theta t} \frac{(\theta t)^j}{j!} \mathbf{K}_h^j d\tilde{H}(t)$$

$$= \sum_{j=n}^{\infty} \gamma_j \mathbf{K}_h^j \tag{D.4}$$

as $\mathbf{K}_h^j = 0$ for $j < n$. In practice, truncation of the sum in Equation D.4 is needed.

When the service rate is deterministic γ_n can be obtained recursively [25] as:

$$\gamma_0 = e^{-\theta d^{-1}} \tag{D.5a}$$

$$\gamma_n = \left(\frac{\theta d^{-1}}{n}\right)\gamma_{n-1} \tag{D.5b}$$

where d denotes the deterministic service rate.

Appendix E:

Computation of Stationary Queue Length Distribution Vector x

The computation of the stationary probability vector **x** at departure epochs required in the analysis of the BMAP/D/1/N queue in Section 4.3.2 of Chapter 4 can be carried out as follows. Recall (Equation 4.15) that:

$$\mathbf{xP} = \mathbf{x}, \quad \mathbf{xe} = 1$$

where $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, .., \mathbf{x}_{N-1})$ giving

$$(\mathbf{x}_0, \mathbf{x}_1, .., \mathbf{x}_{N-1}) = (\mathbf{x}_0, \mathbf{x}_1, .., \mathbf{x}_{N-1})\mathbf{P}, \quad \mathbf{xe} = 1.$$

The vector **x** can be obtained from [6]:

$$\mathbf{x} = (0 \ldots 0\ 1)\ (\mathbf{I} - \mathbf{P}_1)^{-1}$$

where matrix $\mathbf{P}_1$ was obtained from matrix **P** by replacing the last column by $(-1\ -1\ \ldots -1\ 0)^T$. Then $\mathbf{I} - \mathbf{P}_1$ can be written as:

$$\mathbf{I} - \mathbf{P}_1 = \begin{bmatrix} \mathbf{F}_0 & \mathbf{F}_1 & \mathbf{F}_2 & . & . & \mathbf{F}_{N-2} & \mathbf{F}_{N-1} \\ \mathbf{E}_0 & \mathbf{E}_1 & \mathbf{E}_2 & . & . & \mathbf{E}_{N-2} & \mathbf{G}_{N-1} \\ 0 & \mathbf{E}_0 & \mathbf{E}_1 & . & . & \mathbf{E}_{N-3} & \mathbf{G}_{N-1} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & \mathbf{E}_0 & \mathbf{G}_1 \end{bmatrix}$$

By applying the permutation matrix **S** given by:

$$\mathbf{S} = \begin{bmatrix} & & & & & & \\ 0 & \mathbf{I} & 0 & . & . & 0 & 0 \\ 0 & 0 & \mathbf{I} & . & . & 0 & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & 0 & . & . & 0 & \mathbf{I} \\ \mathbf{I} & 0 & 0 & . & . & 0 & 0 \\ & & & & & & \end{bmatrix}$$

we obtain:

$$\mathbf{S}(\mathbf{I}-\mathbf{P}_1) = \left[\begin{array}{ccccc|c} \mathbf{E}_0 & \mathbf{E}_1 & \mathbf{E}_2 & . & \mathbf{E}_{N-2} & \mathbf{G}_{N-1} \\ 0 & \mathbf{E}_0 & \mathbf{E}_1 & . & \mathbf{E}_{N-3} & \mathbf{G}_{N-2} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & . & \mathbf{E}_0 & \mathbf{G}_1 \\ \hline \mathbf{F}_0 & \mathbf{F}_1 & \mathbf{F}_2 & . & \mathbf{F}_{N-2} & \mathbf{F}_{N-1} \end{array}\right] = \begin{bmatrix} \mathbf{T} & \mathbf{D} \\ \mathbf{C} & \mathbf{B} \end{bmatrix} \qquad \text{(E.1)}$$

If the matrix $\mathbf{A}_0^{-1}$ exists then so does $\mathbf{E}_0^{-1}$ and hence $\mathbf{T}^{-1}$ exists.

Now using the Schur-Banachiewicz formula for the inverse of block matrices [45] $(\mathbf{S}(\mathbf{I}-\mathbf{P}_1))^{-1}$ can be obtained as:

$$(\mathbf{S}(\mathbf{I}-\mathbf{P}_1))^{-1} = \begin{bmatrix} \mathbf{T}^{-1}+\mathbf{E}\Delta^{-1}\mathbf{F} & -\mathbf{E}\Delta^{-1} \\ -\Delta^{-1}\mathbf{F} & \Delta^{-1} \end{bmatrix} \qquad \text{(E.2)}$$

where $\Delta = \mathbf{B} - \mathbf{C}\mathbf{T}^{-1}\mathbf{D}$, $\mathbf{E} = \mathbf{T}^{-1}\mathbf{D}$, and $\mathbf{F} = \mathbf{C}\mathbf{T}^{-1}$

The matrix Δ is referred to as the Schur complement of the matrix **T**. **T** is a Toeplitz matrix (i.e. matrix whose right diagonal entries are identical) and its inverse can be obtained in the manner shown in Appendix F. Having calculated $\mathbf{T}^{-1}$, matrix Δ can be obtained next.

The required vector **x** can then be obtained from Equation E.2 after postmultiplying it by **S**. As a result, it can be obtained as the last row of matrix:

$$(\Delta^{-1} \quad -\Delta^{-1}\mathbf{F}). \qquad \text{(E.3)}$$

Appendix F:

Method for Computation of Inverse of Toeplitz Matrix

The computation of the inverse of a Toeplitz matrix is required as a step towards calculation of the stationary queue length distribution at departure epochs x presented in Appendix E.

A Toeplitz matrix is a matrix whose right diagonal entries are identical. Consider a 3×3 Toeplitz matrix T given by:

$$T = \begin{bmatrix} a_0 & a_{-1} & a_{-2} \\ a_1 & a_0 & a_{-1} \\ a_2 & a_1 & a_0 \end{bmatrix}$$

Then T^{-1} can be obtained as [31], [50], [5]:

$$T^{-1} = \begin{bmatrix} x_0 & 0 & 0 \\ x_1 & x_0 & 0 \\ x_2 & x_1 & x_0 \end{bmatrix} \frac{1}{v_0} \begin{bmatrix} v_0 & v_1 & v_2 \\ 0 & v_0 & v_1 \\ 0 & 0 & v_0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ y_0 & 0 & 0 \\ y_1 & y_0 & 0 \end{bmatrix} \frac{1}{y_2} \begin{bmatrix} 0 & u_0 & u_1 \\ 0 & 0 & u_0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \text{(F1)}$$

where

$$x = (x_0, x_1, x_2)$$

$$y = (y_0, y_1, y_2)$$

$$u = (u_0, u_1, u_2)$$

$$v = (v_0, v_1, v_2)$$

are the solutions to the following set of equations:

$$a_0x_0 + a_{-1}x_1 + a_{-2}x_2 = I$$
$$a_1x_0 + a_0x_1 + a_{-1}x_2 = 0$$
$$a_2x_0 + a_1x_1 + a_0x_2 = 0$$

$$a_0y_0 + a_{-1}y_1 + a_{-2}y_2 = 0$$
$$a_1y_0 + a_0y_1 + a_{-1}y_2 = 0$$
$$a_2y_0 + a_1y_1 + a_0y_2 = I$$

$$u_0a_0 + u_1a_1 + u_2a_2 = 0$$
$$u_0a_{-1} + u_1a_0 + u_2a_1 = 0$$
$$u_0a_{-2} + u_1a_{-1} + u_2a_0 = I$$

$$v_0a_0 + v_1a_1 + v_2a_2 = I$$
$$v_0a_{-1} + v_1a_0 + v_2a_1 = 0$$
$$v_0a_{-2} + v_1a_{-1} + v_2a_0 = 0$$

The above expressions apply also in the case of a block Toeplitz matrix where all entries are themselves matrices with the same dimension. Block matrix **T** whose inverse is required in Section 4.4 and in Appendix E is a simple case of a block Toeplitz matrix with $\mathbf{a}_i = 0 \;\; i \geq 1$. This simplifies the above equations removing the last two sets (as $\mathbf{u}_i = 0 \;\; i \geq 0$). Also $\mathbf{x}_i = 0 \;\; i \geq 1$. This results in matrix $\mathbf{T}^{-1}$ being given (for the 3×3 case) as:

$$\mathbf{T}^{-1} = \frac{1}{\mathbf{v}_0}\begin{bmatrix} \mathbf{x}_0 & 0 & 0 \\ 0 & \mathbf{x}_0 & 0 \\ 0 & 0 & \mathbf{x}_0 \end{bmatrix}\begin{bmatrix} \mathbf{I} & \mathbf{v}_0^{-1}\mathbf{v}_1 & \mathbf{v}_0^{-1}\mathbf{v}_2 \\ 0 & \mathbf{I} & \mathbf{v}_0^{-1}\mathbf{v}_1 \\ 0 & 0 & \mathbf{I} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}_0 & \mathbf{t}_{-1} & \mathbf{t}_{-2} \\ 0 & \mathbf{x}_0 & \mathbf{t}_{-1} \\ 0 & 0 & \mathbf{x}_0 \end{bmatrix} \qquad \text{(F2)}$$

The above scheme can be generalised and the resulting algorithm will then look as follows:

1. Calculate $\mathbf{x}_0 = \mathbf{v}_0 = \mathbf{a}_0^{-1}$

2. Calculate $\mathbf{b}_{-i} = \mathbf{a}_{-i}\mathbf{a}_0^{-1}$ for $i \geq 1$

3. Calculate $\mathbf{t}_{-i} = \sum_{m=0}^{i-1} \mathbf{t}_{-m}\mathbf{b}_{m-i}$ for $i \geq 1$

Appendix G:

Use of Supplementary Variable Technique for Queue Length Distribution

The queue length distribution y_i for $1 \leq i \leq N$ required for the solution of the BMAP/D/1/N queue in Section 4.3.2 of Chapter 4 can be found using the supplementary variable technique [11], [36], [30]. The idea here is to consider a pair of variables: one describing the queue length process itself at time t and the second (supplementary) variable representing at time t the time between t and the following departure epoch (i.e. the remaining service time for a customer in service at time t). Remaining service time which is also referred to as forward recurrence service time and the corresponding backward recurrence service time are depicted in Figure G.1.

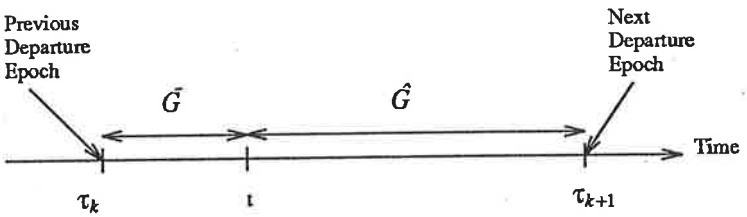


Figure G.1: Forward ($\hat{G}$) and Backward ($\vec{G}$) Recurrence Service Time

Consider the joint distribution of the queue length and the remaining service time for the customer in service given that the server is busy. Define:

$$p_{ij}(\tau)d\tau = P\left\{X(t)=i, J(t)=j, \tau < \hat{G} \le \tau + d\tau \mid \textit{server is busy at } t\right\} \tag{G.1}$$

Define its Laplace transform as:

$$p^*_{ij}(s) = \int_0^\infty e^{-s\tau} p_{ij}(\tau)d\tau \tag{G.2}$$

Note that:

$$p^*_{ij}(0) = \int_0^\infty p_{ij}(\tau)d\tau$$

$$= P\left\{N(t)=n, J(t)=j \mid \textit{server busy at } t\right\}$$

Then

$$\mathbf{y}_i = \mathbf{P}\left\{N(t)=i \mid \textit{the server is busy}\right\} P_{busy} = \mathbf{p}^*_i(0) P_{busy}. \tag{G.3}$$

where $\mathbf{p}^*_i(0)$ is a vector: $(p^*_{i1}(0),..., p^*_{im}(0))$

The integral in Equation G.2 can be evaluated by conditioning on the number of arrivals occurring during the elapsed service time $\vec{G}$ for the customer in service. Let

$$h_{j'j}(i)(\tau)d\tau = P\left\{i \textit{ arrivals in } \vec{G}, J(\tau_k + \vec{G}) = j, \tau < \vec{G} \le \tau + d\tau \mid J(\tau_k) = j'\right\} \tag{G.4}$$

with its Laplace transform as:

$$H_{j'j}(i)(s) = \int_0^\infty e^{-s\tau} h_{j'j}(i)(\tau)d\tau \tag{G.5}$$

and let $\mathbf{H}_i(s)$ be a matrix with its (j,j')th elements given by $H_{j'j}(i)(s)$. Then $\mathbf{p}^*_i(s)$ for $1 \le i \le N-1$ can be obtained as [6]:

$$\mathbf{p}^*_i(s) = \mathbf{x}_0 \sum_{\nu=1}^{i} \mathbf{U}_\nu(0)\mathbf{H}_{i-\nu}(s) \quad (\textit{queue empty})$$

$$+ \sum_{\nu=1}^{i} \mathbf{x}_\nu \mathbf{H}_{i-\nu}(s) \quad (\nu \textit{ customers in queue}) \tag{G.6}$$

with $\mathbf{U}_\nu(0)$ as defined Appendix C.

After extensive algebraic manipulation [6] and [32] $\mathbf{p}^*{}_i(s)$ can be reduced to:

$$\mathbf{p}^*{}_i(s) = \mu\left[\sum_{\nu=0}^{i-1}\mathbf{x}_\nu\ \mathbf{D}_{i-\nu-1}(s) + \mathbf{x}_0 H(s)\sum_{\nu=1}^{i}\mathbf{D}_0^{-1}\mathbf{D}_\nu\,\mathbf{D}_{i-\nu}(s) - \sum_{\nu=1}^{i}\mathbf{x}_\nu H(s)\mathbf{D}_{i-\nu}(s)\right] \qquad \text{(G.7)}$$

for $1 \le i \le N-1$ and where $\mathbf{D}_n(s)$ can be obtained as the coefficients of z^n in the expansion of :

$$[\mathbf{D}(z) + s\,\mathbf{I}]^{-1} = \sum_{n=0}^{\infty}\mathbf{D}_n(s)z^n.$$

$H(s)$ is the Laplace-Stieltjes transform of the service distribution $\bar{H}(t)$ given by:

$$H(s) = \int_0^\infty e^{-st}\,d\bar{H}(t)$$

Substituting Equation 4.24 (which gives the expression for P_{busy}) and Equation G.6 into Equation G.3, and noting that $H(0) = 0$ gives:

$$\mathbf{y}_n = \frac{1}{\mu^{-1} - \mathbf{x}_0\mathbf{D}_0^{-1}\mathbf{e}}\left[\mathbf{x}_0[\mathbf{D}_{n-1}(0) + \sum_{\nu=1}^{n}(\mathbf{D}_0^{-1}\mathbf{D}_\nu)\mathbf{D}_{n-\nu}(0)]\right.$$

$$\left. + \sum_{\nu=1}^{n-1}\mathbf{x}_0[\mathbf{D}_{n-\nu-1}(0) - \mathbf{D}_{n-\nu}(0)] - \mathbf{x}_n\,\mathbf{D}_0(0)\right] \qquad \text{(G.8)}$$

for $1 \le n \le N-1$.

Appendix H:

Obtaining U_D from U_C in Superposition Method II

The expressions for obtaining the set of defining parameters $U_D = (\lambda_1, \lambda_2, r_1, r_2)$ for a two-state MMPP from the set of characterising parameters $U_C = (\lambda, c^2, Z(\infty), C(\infty)$ are presented below. The analysis used in developing them is reproduced here from [51] using a somewhat different notation.

The mean arrival rate λ and the coefficient of variation of interarrival times c^2 are given by:

$$\lambda = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}, \quad \lambda_1 > \lambda_2, \quad \lambda > 0 \tag{H.1}$$

and

$$c^2 = 1 + \frac{2 r_2 r_1 (\lambda_1 - \lambda_2)^2}{(\lambda_1 r_2 + \lambda_2 r_1 + \lambda_1 \lambda_2)(r_1 + r_2)^2}$$

$$= 2\lambda \frac{\lambda_1 + \lambda_2 + r_1 + r_2 - \lambda}{\lambda_1 r_2 + \lambda_2 r_1 + \lambda_1 \lambda_2} - 1, \quad c^2 > 1 \tag{H.2}$$

Introduce an intermediate set $U = (\lambda, c^2, b, \nu)$ where λ and c^2 are as given above and

$$b = \frac{Z(\infty) - c^2}{Z(\infty) - 1} = \frac{\lambda_1 \lambda_2}{\lambda_1 r_2 + \lambda_2 r_1 + \lambda_1 \lambda_2}, \quad b < 1 \tag{H.3}$$

$$\nu = \frac{\lambda^2 (Z(\infty) - 1)^2}{4C(\infty)} = \frac{r_1 r_2 (\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2}, \quad \nu > 0 \tag{H.4}$$

Now let

$$k_1 = r_1 + r_2 + \lambda_1 \tag{H.5}$$

$$k_2 = r_1 + r_2 + \lambda_2 \tag{H.6}$$

$$p_1 = \frac{1}{2}[(\lambda_1 + \lambda_2) + r_1 + r_2 - \Delta] \tag{H.7}$$

$$p_2 = \frac{1}{2}[(\lambda_1 + \lambda_2) + r_1 + r_2 + \Delta] \tag{H.8}$$

where

$$\Delta = [(\lambda_1 - \lambda_2) + r_1 - r_1]^2 + 4r_1r_2$$

then

$$\lambda_1 = p_1 + p_2 - k_2 \tag{H.9}$$

$$\lambda_2 = p_1 + p_2 - k_1 \tag{H.10}$$

$$r_1 = \frac{p_1p_2 - k_2\lambda_1}{k_2 - k_1} \tag{H.11}$$

$$r_2 = \frac{p_1p_2 - k_1\lambda_2}{k_1 - k_2} \tag{H.12}$$

Putting expressions given in Equations H.9-H.12 into Equations H.1-H.4 we get:

$$\lambda = \frac{p_1p_2(1-b)}{k - \lambda} \tag{H.13}$$

$$c^2 = 2\lambda\frac{p_1 + p_2 - \lambda}{p_1p_2} - 1 \tag{H.14}$$

$$b = \frac{(k - \lambda - k_1)(k - \lambda - k_2)}{p_1p_2} \tag{H.15}$$

$$v = -(k - k_1)(k - k_2) \tag{H.16}$$

where

$$k = r_1 + r_2 + \lambda = \frac{2v(1-b)}{\lambda(c^2-1)} + \lambda$$

From Equations H.13, H.14 and the above expression for k, we have:

$$p_1 p_2 = \frac{2v}{c^2-1} \tag{H.17}$$

$$p_1 + p_2 = \frac{v(c^2+1)}{\lambda(c^2-1)} + \lambda \tag{H.18}$$

Combining Equations H.15, H.16 and H.17 and solving the resultant quadratic equations for k_1 and k_2 gives:

$$k_1 = k - \frac{1}{2}[\lambda - \frac{2vb}{\lambda(c^2-1)} - \frac{v}{\lambda} - \sqrt{\delta}] \tag{H.19}$$

$$k_2 = k - \frac{1}{2}[\lambda - \frac{2vb}{\lambda(c^2-1)} - \frac{v}{\lambda} + \sqrt{\delta}] \tag{H.20}$$

where

$$\delta = [\lambda - \frac{2vb}{\lambda(c^2-1)} - \frac{v}{\lambda}]^2 + 4v$$

The expressions given by Equations H.17 to H.20 allow calculation of U_D from the intermediate set U (using Equations H.9 to H.12) and hence from U_C (using Equations H.3 and H.4).

References

1. Albin S.,"Approximating a Point Process by a Renewal Process II: Superposition Arrival Processes to Queues", Operations Research, vol. 32, 1133-1162, September-October 1984.

2. Anido G., "Traffic Control and Management Mechanism for a Broadband Packet Network", ITC Specialist Seminar, Adelaide, 1989.

3. Appleton J., "Modelling a Connection Acceptance Strategy for Asynchronous Transfer Mode Networks", ITC Specialist Seminar, New Jersey, 1990.

4. Aumann G., "Source Policing in the Broadband-ISDN", 4th ATERB FPS Workshop, Sydney, July 1989.

5. Ben-Artzi A. and Shalom T., "On Inversion of Block Toeplitz Matrices", Integral Equations and Operator Theory, vol. 8, 751-779 (1985).

6. Blondia C., "The N/G/1 Finite Capacity Queue", Commun. Statist.-Stochastic Models, 5(2), 273-294 (1989).

7. Bonomi F., Fratta L., Montagna S., and Paglino R. "Priority on Cell Service and on Cell Loss in ATM switching", ITC Specialist Seminar, New Jersey, 1990.

8. Burgin J. "Resource Management and control in the Broadband Integrated Services Digital Network", PhD Thesis, Monash University, January 1990.

9. CCITT Study Group XVIII - Report R 34, June 1990.

10. Cinlar E., "Markov Renewal Theory", Adv. Appl. Prob., 1, 123-187, 1969.

11. Cohen J.W., The single server queue, North-Holland Series in Applied Mathematics and Mechanics, 1982. (Chapter 6)

12. Cost 224 "On Admission Control and policing in an ATM Based Network", ITC Specialist Seminar, New Jersey, 1990.

13. Doshi B.T., Eckberg A.E., Saksena V.R., and Zoccolillo R., "A B-ISDN/ATM Network Congestion Control Architecture, and Its Performance with Complementary End-Terminal Flow and Error Controls", ITC Specialist Seminar, New Jersey, 1990.

14. Dziong Z., Choquette J., Liao K., and Mason L., "Admission Control and Routing in ATM Networks", ITC Specialist Seminar, Adelaide, 1989.

15. Eckberg A.E. "Generelized Peakedness of Teletraffic Processes", ITC 10, Montreal, 1983.

16. Eckberg A.E., Luan D.T., and Lucantoni D.M., "Bandwidth Management: A Congestion Control Strategy for Broadband Packet Networks - Characterizing the Throughput-Burstiness Filter", ITC Specialist Seminar, Adelaide, 1989.

17. Evans S.P., "A Mathematical Model and Related Problems of Optimal Management and design in a Broadband Integrated Services Network", J. Austral. Math. Soc. ser. B 31(1989), 150-175.

18. Filipiak J. "Flexible Traffic Control Schemes in a Fast Packet Switching Networks", 3rd ATERB FPS Workshop, Melbourne, May 1988.

19. Grassmann W.K., "Finding Transient Solution in Markovian Event Systems Through Randomization", The First International Conference on the Numerical Solution of Markov Chains, North Carolina, 1990, 375-395.

20. Griffiths T.R., "Analysis of Connection Acceptance in Asynchronous Transfer Mode Networks", British Telecom Research Laboratories, 1990.

21. Grimmett G.R. and Stirzaker D.R., Probability and Random Processes, Oxford University Press, 1982.

22. Guillemin F. and Tranchier D., "Supervision dans un Reseau RNIS-Large Bande: La Fonction de Policing", Memoir de Fin d'Etudes, Telecom Paris.

23. Gun L., "Experimental Results on Matrix-Analytical Solution Techniques - Extensions and Comparisons", Commun. Statist.-Stochastic Models, 5(4), 669-682 (1989).

24. Heffes H., "A Class of Data Traffic Processes - Covariance Function Characterization and Related Queueing Results", BSTJ, vol. 59, no. 6, July-August 1979.

25. Heffes H., Lucantoni D.M., "A Markov Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance", IEEE Journal on Selected Areas in Communications, vol. SAC-4, no.6, September 1986.

26. Holtzman J.M. "Characteristics of Superposition of Traffic Streams", AT&T Bell Laboratories, 1989.

27. Hughes D., "Congestion Control in the Broadband ISDN", Report, University of Wollongong, 1990.

28. Hughes D., Anido G., and Bradlow H. "Characterising Leaky Bucket Performance for Small Bucket Depth", 5th ATERB FPS Workshop, Melbourne, July 1990.

29. Hui J.Y. "Resource Allocation for Broadband Networks", IEEE Journal on Selected Areas in Communications, vol. 6, no. 9, December 1988.

30. Ide I. "Superposition of Interrupted Poisson processes and Its Application to packetized Voice Multiplexers", ITC 12 Torino, 1988.

31. Iohvidov I.S., Hankel and Toeplitz Matrices and Forms - Algebraic Theory, Birkhausen Boston, 1982.

32. Jacobsen S.B., Moth K., and Dittman L., "Load Control in ATM Networks", XIII International Switching Symposium, Stockholm, May 1990.

33. Joos P., Verbiest W., "A Statistical Bandwidth Allocation and Usage Monitoring Algorithm for ATM Networks", ICC/89, Boston, June 1989.

34. Kawashima K., and Saito H., "Teletraffic issues in ATM networks", ITC Specialist Seminar, Adelaide, 1989.

35. Le Boudec J.Y., private communication.

36. Lee T.T. ""M/G/1/N Queue with Vacation Time and Exhaustive Service Discipline", Oper. Res. 32, No. 4, 1984, pp. 774-784.

37. Lucantoni D.M., "Efficient Algorithms for Solving the Non-linear Matrix Equations Arising in Phase Type Queues", Commun. Statist-Stochastic Models, 1(1), 29-51 (1985).

38. Lucantoni D.M., "New Results on the Single Server Queue with a Batch Markovian Arrival Process", Stochastic Models, vol. 7, no. 1, 1991.

39. Malgris B., et al, "Performance Models of Statistical Multiplexing in Packet Video Communications", IEEE Transactions on Communications, vol. 36, no. 7, July 1988.

40. Mase K., and Shioda S., "Real-Time Network Management for ATM Networks", to be presented at the 13th ITC, Copenhagen 1991.

41. Meier-Hellstern K.S., "A Fitting Algorithm for Markov-modulated Poisson Processes Having Two Arrival Rates", European Journal of Operational Research 29 (1987), 370-377.

42. Neuts M.F., "A versatile Markovian point process", J. Appl. Prob. vol. 16, pp. 764-779, Dec 79.

43. Neuts M.F., "The Markovian Arrival Process: A Tool for Addressing Broadband Traffic Issues", ITC Specialist Seminar, New Jersey, 1990.

44. Ohnishi H, Okada T, and Noguchi K. "Flow Control Schemes and Delay/Loss Tradeoff in ATM Networks", IEEE J. on Sel. Areas in Com., vol. 6, no. 9, 1988.

45. Ouellette D.V. "Shur Complements and Statistics", Linear Algebra and its Applications vol. 36, pp. 187-295, Mar 1981.

46. Ramaswami V., "The N/G/1 Queue and its Detailed Analysis", Adv. Appl. Prob., vol 12, pp. 222-261, Mar 80.

47. Ramaswami V., "An Experimental Evaluation of the Matrix-Geometric Method for the GI/PH/1 Queue", Commun. Statist.-Stochastic Models, 5(4), 629-667 (1989).

48. Rathgeb E.P. "Policing Mechanisms for ATM Networks - Modelling and Performance Comparison", ITC Specialist Seminar, New Jersey, 1990.

49. Roberts J., and Simonian A.,"Some Queueing Models for an ATM Multiservice Network", Doc. COST 224.

50. Roebuck P.A. and Barnett S., "A Survey of Toeplitz and Related Matrices", Int. J. Systems Sci., 1978, vol.9, No.8, 921-934.

51. Rossiter M.H., "Sojourn Time Theory and the Switched Poisson Process", Telecom Australia Research Laboratories Report 7835, 1986.

52. Rossiter, M.H. "A switched Poisson model for data traffic", Aust. Telecomm. Res., vol 21, 53-57, 1987.

53. Rossiter, M.H., "Characterizing Bursty Traffic for the Purpose of Network Design", 3rd Australian Teletraffic Research Seminar, Melbourne Nov. 1988.

54. Rossiter M.H., "A Survey of Some Recent Results in the Modelling of Bursty Traffic", 4th Australian Teletraffic Research Seminar, Bond University, Dec. 1989.

55. Sallberg K, and Stavenow B., "A Resource Allocation Framework in B-ISDN", XIII International Switching Symposium, Stockholm, 1990.

56. Sriram K., Whitt W., "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data", IEEE Journal on Selected Areas in Communications, vol. SAC-4, no. 6, September 1986.

57. Tran-Gia P., A Class of Renewal Interrupted Poisson Processes and Applications to Queueing Systems", Zeitschrift fur Operations Research, Vol 32, pg. 231-250.

58. Verbiest W., Pinnoo L., and Voeten B. "Statistical Multiplexing of Variable Bit Rate Video Sources in Asynchronous Transfer Mode Networks", Globcom 1988.

59. Wallmeier E., "A Connection Acceptance Algorithm for ATM Network Based on Mean and Peak Bitrates", submitted to Int. J. Digital and Analogue Cabled Systems.

60. Whitt W., "Approximating a Point process by a Renewal Process, I: Two Basic Methods," Oper. Res., 30, no.1 (January-February 1982), pp. 125-147.

61. Woodruff G.M., Rogers R.G.H., and Richards P.S., "A Congestion Control Framework for High-Speed Integrated Packetized Transport, IEEE

GLOBECOM'88, November, 1988.

62. Woodruff G.M. and Kositpaiboon "Multimedia Traffic Management Principles for Guaranteed ATM Network Performance", IEEE J. on Sel. Areas in Comm., Vol. 8, No. 3, April 1990.

63. Zukerman M., "Applications of Matrix Geometric Solutions for Queueing Performance Evaluation of a Hybrid Switching System", J. Austral. Math. Soc. Ser. B 31(1989), 219-239.