# MMPP modeling of ATM multimedia Traffic

**Shahram Shah-Heydari**

**A Thesis**

**in**

**The Department**

**of**

**Electrical and Computer Engineering**

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science at
Concordia University
Montreal, Quebec, Canada

**July 1998**

0-612-39477-8

# ABSTRACT

MMPP Modeling of ATM Multimedia Traffic

Shahram Shah-Heydari

Traffic control in broadband networks has been a topic of interest in many research projects across the world recently. Due to the variety of the offered services in broadband networks like multimedia and video, the traffic flowing in the broadband networks is considered highly bursty. Therefore the traffic control will be very important in preventing congestion, reducing the probability of cell loss, and defining suitable algorithms for call admission.

Traffic control cannot be done without having a well-fit model to represent the traffic. Therefore traffic modeling is an essential part of any network control research project. A number of models have been proposed to represent various types of traffic in broadband networks. Among them, Markov-Modulated Poisson Process (MMPP) shows the great flexibility and analytical tractibility which is needed in traffic control. MMPP model is not only capable of capturing the interframe correlation in the traffic, but also can be easily analysed by using well-known Matrix Geometric techniques.

Our research in this project is focused on the study of MMPP for modeling of the traffic in the broadband networks. We first start with the simplest case, a two state MMPP, and study its performance for representing the ATM traffic. Starting with a superposition of voice sources, the performance of various techniques to model the superposed stream by a 2-state MMPP is compared. Then the techniques are generalized for an arbitrary aggregated ATM traffic, characterized only from a sequence of traffic samples. A refined moment-based technique to derive the parameters of a 2-state MMPP model to represent such an arbitrary traffic is proposed. Our simulation results show a high degree of accuracy in parameter estimation. We also present an approximation for the probability of loss in a 2-state MMPP/D/1 queue. Therefore, based on the measured data from an ATM traffic source, we can use the proposed technique to model the traffic source by a 2-state MMPP and then apply the approximation to predict its probability of loss.

There are some cases where two states are not enough for representing the change of the phases in the traffic. In order to have a more general model applicable to various types of traffic, we propose a special type of multiple-state MMPP, a superposition of N 2-state MMPP minisources. This model, besides simplicity, enjoys all the advantages of MMPP models. Its parameters can be found from empirical data. We propose a pdf-based technique to derive the parameters of the model from the traffic samples. Using several examples as well as some case studies we show the accuracy of the technique in parameter estimation and its power to represent ATM traffic. An approximation for the slope of the curve of the probability of loss versus buffer size is also derived.

**Keywords:** ATM, Traffic Modeling, MMPP, Markov-Modulated Poisson process, Multimedia Traffic, Pdf-Based Matching.

*To my dear wife Mahnaz,*

*For her patience, love and support all the time.*

# ACKNOWLEDGMENTS

I would like to express my sincere thanks to my thesis supervisor, Prof. Tho Le-Ngoc for his invaluable guidance and support without which it would have been impossible to complete this thesis.

I gratefully acknowledge the role of the Canadian Institute for Telecommunications Research (CITR) in funding this project.

I would like to thank my colleagues and friends at Concordia University, especially Ehsan Rezaiifar, Hamid Mehrvar, T.V.J Ganesh Babu, Tien Hy Bui, Periklis Tsingotjidis, Rajab Faraj and Rocco DiGirolomo for their support and encouragement.

Finally, I would also like to extend my gratitude to faculty and staff members of the department of Electrical and Computer Engineering, Concordia University.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1
# Introduction

## 1.1 Background

The emergence of Internet as the main future communication medium around the globe has dramatically enhanced the role of telecommunication technology in the world. Now the communication networks are not supposed to carry voice telephony only, the role they performed throughout the 20th century. The start of the new millenium witnesses the great technical advances which promise future multimedia services including -but not limited to- voice telephony, video-on-demand, teleconferencing and videoconferencing, various data services, from low traffic, transaction-based services such as banking on the net and email, to highly bursty traffic with huge volume of data services such as web browsing and file transfer, and most important, all of these services must be carried on a unified, broadband network.

Such requirements, of course, need a lot of research and study in network design. The designers must take several factors into account, most important of them is the *Quality of Service* (QoS) [7] which must guarrantee a certain degree of performance for the end user. This requires mathematical methods for the analysis of the networks. Furthermore, such dynamic networks with a traffic which changes its pattern all the time cannot be controlled in a static way. The traditional public switched telephony networks had a simple, circuitswitched network whose traffic could be modelled using Poisson process. The advances in the queueing theory in the second half of the 20th century made the analysis and design of such networks straightforward. Such assumptions are not valid anymore for the current multimedia networks. The circuit-switching, though makes it easy to guarrantee the lowest possible delay and the best performance, wastes such a huge bandwidth that the broadband networks simply cannot afford. Therefore it has been replaced by the more efficient packet-switched approach. The high-level protocol is also changed to internet protocol TCP/IP. The emergence of Asyncronous Transfer Mode (ATM) technology has changed many issues in the design too. ATM technology uses a fixed-length cell throughout the network. All of the above points indicate the need for a comprehensive network design approach which first of all requires a reasonably accurate, feasible network performance evaluation. Such evaluation must enable us not only to analyse and predict the performance of the network at the design stage, but also to predict it on line, in order to activate appropriate network control and call admission procedures.

Traffic modeling is one of the main topics in network performance evaluation. Appropriate traffic models can be used in the simulators as traffic generator, or be used in deriving performance indicators for the purposes of network resource allocation and control in real time in dynamic networks [7]. Buffer and switch designs largely depend on the profile of the traffic, therefore, the traffic models must be capable of capturing the most important characteristics of the traffic which affect the network performance.

Network traffic is a stochastic process. It means that the intervals between the incoming cells or packets are random variables because of the uncertainty in the behaviour of the end users. Therefore the models proposed for network traffic must be probabilistic too. One of the first proposed model was the simple Poisson Process, first introduced in 19th century and was found to be a good approximation for receiving calls in a telephony network. However, the evolution in the telecommunication networks and the increase in the complexity as well as introducing various type of services, make Poisson Process a poor choice for network traffic modeling [7]. The most important shortcoming of Poisson process is its memorylessness, or independency of the interarrival times. This means that the Poisson process is unable to capture the correlation between the consecutive frames. In reality usually such correlation exists. When a user starts downloading a file from Internet, usually it receives traffic for a specific time interval which varies depending on the network load. Even during phone calls, usually one talks for a few seconds and listen for another few seconds. This shows a certain pattern in the traffic and nullifies the independency assumption. Even more, in recent years several studies have shown traces of the self-similarity phenomenon in the Internet traffic [8]. The self-similarity property indicates very long-range dependency in the traffic which repeats itself over various time intervals from miliseconds up to hours and days. It further complicates the traffic modeling in ATM multimedia networks.

Several models have so far been introduced for various types of traffic. Among them, the Markov-Modulated Poisson Process (MMPP) shows the maximum flexibility required for traffic modeling [7]. Using an unlimited choice of the number of phases, MMPP in the simplest 2-state case or in multiple-state case has been shown to be capable of modeling various types of traffic such as aggregated voice channels and video sources. While MMPP is not a self-similar model, several studies suggest that it may be used for modeling of long-range dependent data traffic under certain assumptions of traffic control. The main attractiveness of MMPP, besides its flexibility, lays in the analytical tractability of the MMPP model. MMPP/G/1 queues have been discussed and analysed for almost two decades and well known analytical algorithms such as Matrix Geometric technique are available for analysing the systems with this type of the traffic as the input [18].

## 1.2 Research objectives

Our main concern in this research is how to model an ATM traffic by an MMPP model. The problem can be described in simple terms: Suppose we have a given traffic stream of ATM cells, and we want to find the parameters of the model in a way that it can represent the corresponding traffic well. In other words, the model must be able to be used for the prediction of the performance of the traffic under various conditions (traffic loads, buffer sizes, etc).

The given traffic stream is a sequence of samples extracted from the real data. The samples may show the time intervals between cell arrivals, and in this case are called the *Time* process, or may show the number of arrivals over a fixed observation interval, and in this case it is called the *counting* process. Both processes are mathematically equivalent. However, in this thesis, we have limited our research to the counting process which is easier to simulate and to work with.

We also divide our modeling task to two separate areas, one for the simplest case, 2-state MMPP, and the other one for multiple-state MMPP. The reason is that the simplicity of the 2-state MMPP enables us to use some simpler and more accurate techniques which are inapplicable in a general multiple-state MMPP case. Nevertheless, the multiple-state case is studied too because there are cases where a two-state model cannot capture various phases which exist in the traffic.

In order to assess the performance of the model, we usually form two separate systems, one a G/D/1 queue for the sample traffic stream and another an MMPP/D/1 queue for our derived model. We assume that if the performance of both systems (including cell delay and probability of loss under various traffic loads) are the same or close, then the model is acceptable. Here we use *OPNET* network simulator [34] to simulate these queueing systems. The results of the simulation have been analysed by *MATLAB* software tool.

### 1.3 Scope of the thesis

In Chapter 2 we introduce the basis for traffic modeling and characterizations. Various traffic indicators are discussed and several models for video, voice and data traffic are reviewed. A separate section has been dedicated to study the MMPP model.

Chapter 3 is dedicated to the 2-state MMPP model. We start the study of MMPP modeling with the simplest case, the aggregated voice traffic. Several techniques for matching a 2-state MMPP model to an aggregated voice traffic are discussed, and their performances are analysed and compared by using simulation. Then the matching technique is generalized for a general, arbitrary traffic rather than an aggregate voice traffic. The performance of the model is studied and also an approximation of the probability of loss in the 2-state MMPP/D/1 queue is derived.

In Chapter 4 a more general model, a special case of multiple-state MMPP is introduced for modeling ATM multimedia traffic. A new pdf-based technique is proposed to derive the parameters of the model. An approximation for the slope of the curve of the probability of loss is also presented.

Finally, in Chapter 5, the conclusions of the work and some suggestions for future work are presented.

# CHAPTER 2
# Traffic Characteristics and Modeling

## 2.1 Introduction

In this chapter we start with the main characteristics of ATM network traffic. We examine some parameters of the traffic which are more important in the queueing behaviour. Then we will survey various models proposed for modeling of the traffic sources, for different types of traffic (voice, video and data). We study our selected model, MMPP (Markov-Modulated Poisson Process) in detail.

## 2.2 Traffic source characteristics

Source characterization defines the parameters of the traffic source which are important in the study of the network behaviour with that traffic. These parameters can be used for source modeling. They are also used by the network management system to allocate its resources among different users, in order to avoid congestion and define and maintain a measure for Quality of Service (QOS) which is negotiated at the time of the connection. They are also used to determine whether to accept a call into the network or not (Call Admission Control). According to CCITT, the following parameters are important in source characterization [3]:

- *Peak Arrival Rate*: The maximum cell arrival rate or the maximum amount of network resources requested by the source. This parameter may alternatively be defined as the reciprocal of the minimum interarrival time between two consecutive cells belonging to the same connection. It is sometimes called *Instantaneous Peak Cell Rate* too.

- *Average Arrival Rate*: The average cell arrival rate or the average amount of network resources requested by the source. It may be the *True Average Cell Rate*, the total number of cells generated during a connection divided by the elapsed time, or the *Estimated Average Cell Rate*, the estimation of the true average over a long time interval T.

- *Burstiness*: The burstiness can be viewed as a measure of the duration of the activity period of a connection. One of the widely used definition for burstiness is the ratio of the peak cell rate to the average cell rate.

- *Burst length*: The average duration of the active state.

There are some other measures who help in characterizing the traffic and modeling it in an efficient way. In the next section we will discuss them in more details.

### 2.3 Performance measures for Traffic modeling

When we try to model the ATM traffic, our ultimate goal is to come up with a mathematical description that can provide us with a prediction of the queueing performance of the network. Before getting into the details of various models for each type of traffic, we must know our criteria for deciding whether a specific model is performing satisfactory or not. Of course in the ideal case, we prefer a model that behaves *exactly* in the same way the original traffic would behave under all conditions. However, one must note that due to the stochastic nature of the network traffic, it is almost impossible even to characterize the real physical traffic, let alone finding an exact mathematical model for it, which is already difficult enough even for deterministic physical processes. Therefore, like any other case, we try to find a model which can predict those performance measures which are the most important to us.

It is generally accepted that the following measures form a rather good criteria for deciding whether a model performs well or not [7]. A model which predicts these measures well enough for a given traffic and system, is accepted.

- *Quality of Service (QoS) Parameters :* These are some general parameters which show the performance of the system. Every model must be able to predict the QoS for the traffic which it is trying to model. Therefore these parameters may be used to evaluate each model. The main QoS parameters are as follows:

  - *Average Delay :* In general the end-to-end delay, i.e., the average time it takes for a cell to reach from input to the system to the output of the system, is a good measure of performance. This includes all queueing delays at the buffers, transmission delays, switching delays and propagation delays. In most cases the first one, the queueing delay, is the most dominant. Transmission delays and switching delays are considered fixed (constant) in ATM networks, because the cell length is fixed. Propagartion delay, although dominant in some cases such as satellite networks, but still is considered constant in ATM networks and does not play any role in network stochastic analysis. Therefore in the analysis of ATM networks, these types of delay are either ignored or added to the total delay as a constant. Average delay can be replaced by the queueing delay as a performance criteria in most of the cases. Based on *Little*'s formula [11], average queue length can also replace the queueing delay. Average delay is the most important parameter in real-time applications, such as voice and video traffic transmission.

- *Probability of Cell Loss :* In case that the buffer has a finite capacity, the extra cells may be discarded or *lost.* Also in some types of switches, such as *Knockout* switch, some of the cells who are destined for the same output port may be discarded. Probability of Cell Loss is the most important factor for computer data traffic (such as internet) because losing a cell forces the sender to re-transmit it based on the communication protocol. We will explain the difference between real-time and non real-time traffic in more details later.

In order to design a buffer, one wants to know what size of buffer must be chosen for a required probability of loss (or probability of *buffer overflow*, in some texts). It is very time-consuming to do analysis or run simulation for various buffer sizes and then come up with a curve which shows the probability of loss versus buffer size for the design purposes. Therefore network designers most of the time use an approximation to simplify the job. Instead of assuming a finite buffer and changing its size and calculating the probability of loss, it is preferred to solve the problem for an infinite buffer, and compute the probability density function of the queue length in this case. Then the survivor function of the pdf of the queue length, Pr(queue length > X), indicates the probability that the queue size goes beyond a specific length for an infinite buffer. Then this probability is used instead of the probability of cell loss. Of course there is an approximation here. The behaviour of large buffers is *assumed* to be identical to that of the infinite one. However, this assumption is not so off when the buffer size is large and it simplifies the analysis and/or the simulation noticably. Also, it is not difficult to show that mathematically the value of the average queue length can be calculated from this survivor function. Therefore, the survivor function of queue length is a good measure to replace both the probability of loss and average queue length (or delay). In this work wherever we point to the *probability of cell loss*, we mean this approximated form, the *survivor function* of the queue length for infinite buffer case.

- *Index of Dispersion for Counts (IDC):* If we denote the number of arrivals over a time interval of t by random variable X(t), then the Index of Dispersion for Counts is defined as the ratio of the variance of X(t) over the mean of X(t). By computing this parameter for different values of time interval t, we will have a curve for IDC(t) versus t. Although IDC curve is a measure of characterization of the traffic rather than a measure of queueing behaviour, it has been shown that this curve has a definite effect on the queueing performance [25]. Any model must have an IDC curve as close as possible to the origical traffic, to have the same queueing performance as it. We will show the effect of IDC curve on the queueing performance in the next chapter. The advantage of IDC curve is that it is computed from the traffic itself, not from its queueing behaviour. So even without any simulation or analysis of the queue, IDC is computable from the input traffic itself. It saves us from unnecessary extra work by looking at IDC curve of the model first before analysing the queue using it.

In our work, we mainly used IDC and survivor function of the queue length as our measures for the performance of the models.

## 2.4 ATM traffic modeling

Several surveys have been done in the past on the subject of ATM traffic modeling. The reader may refer to [7], [3], [2], [4], [1], [5] or an excellent chapter in [10].

ATM traffic can be categorized from the service point of view. The following services are available in ATM:

- *Constant Bit Rate (CBR) :* Here a fixed part of the bandwidth is allocated to the connection. Mainly for continuous, fixed rate bit streams which cannot tolerate delay or jitter, such as voice. The traffic pattern is deterministic.
- *Variable Bit Rate (VBR) :* Defined with two parameters: average bit rate and peak bit rate. For those types of traffic who need minimal cell delay variation and have a bursty traffic pattern, such as video.
- *Available Bit Rate (ABR) :* A minimum bandwidth is guaranteed, and over that up to the current available bandwidth is allowed. Suitable for bursty, delay tolerant traffic such as LAN data.
- *Unspecified Bit Rate (UBR) :* No guaranteed quality of service.

From another point of view, ATM multimedia traffic may be categorized into real-time and non real-time (or jitter tolerant) traffic. The term *jitter* points to the cell delay variation. Obviously for CBR such a variation does not happen. Real time services such as voice and video do not tolerate jitter, so CBR or VBR services must be used for them. For data the variation in delay is not important. Therefore ABR and UBR can be used for data traffic.

Now let us examine briefly various types of ATM mutimedia traffic and the models proposed for each of them.

## 2.5 Voice Models

Voice services have been and continue to be an important part of any broadband communication service. The properties of voice traffic depend upon the encoding scheme adopted. The coding scheme may be fixed rate (such as PCM, DPCM, DPCM or CELP) or variable rate [5]. However fixed rate coding schemes are those which are widely used. For voice services the main important QoS parameter is delay, which is not tolerated. However, as we discussed before, a certain level of the probability of loss is acceptable.

The traffic from an interactive voice source looks like a cell stream modulated by arrivals during talkspurts and no arrivals during silences. For this reason, an on-off model looks like natural for voice sources.

### 2.5.1 On-Off Model

On-off model is a two-state markov model which alternates between phases of activity and silence phases (Figure 2.1). During active phases, cells are generated at a fixed rate. During silence phases, no cells are generated. The sojourn time at each state (the time length of each phase) is a random variable with an exponential form probability density function. The assumption of exponential distribution for talkspurt phase (state ON) is in agreement with the measurement, but for silent period it is not a perfect fit [30]. Nevertheless, in the analysis of On-off sources, the probability density function of the length of the silent phase can be chosen arbitrarily [17]. The On-Off model can be approximated as a renewal process [30].

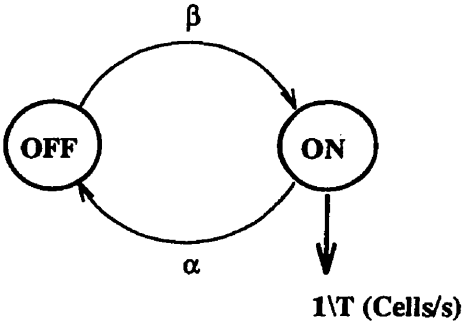

Figure 2.1: On-Off model

As Figure 2.1 indicates, the on-off model is fully described by three parameters: $\alpha$ denotes the mean transition rate from state ON to state OFF, $\beta$ denotes the mean transition rate from state OFF to state ON, and T denotes the constatnt interarrival time in state ON. The typical values of the parameters for PCM voice source is $\alpha^{-1}$ = 352ms and $\beta^{-1}$ = 650ms [17].

In [17] many of the characteristics of on-off model, including the probability density function of interarrival time, the moments of the counting process and the value of index of dispersion for counts (IDC) for large lags have been derived. In the same reference it has been shown that a superposition of statistically multiplexed voice sources can be modeled by a Markov-Modulated Poisson Process (MMPP) which we will describe in the next chapter.

One of the advantages of on-off model, besides its simplicity, is its analytical tractibility. On-off/D/1 queues can be solved analytically by using fluid flow approximation technique [10].

### 2.5.2 IPP model

IPP stands for *Interrupted Poisson Process.* This model is slightly different from on-off in the way that in IPP model, the cell generation in state ON is governed by a Poisson process rather than a deterministic constant rate. So the parameter T here denotes the mean interarrival time which is an exponentially distributed random variable. Although on-off model seems more relevant for voice traffic, but IPP can be used for variable rate coded voice sources. IPP is a special case of two-state MMPP with mean cell generation rate of zero in one of the states. So it is analytically tractable in the same way as MMPP is (Section 2.8).

## 2.6 Video Models

It is expected that video will be a major source on future broadband ATM networks because of all of the multimedia services. Applications such as video conferencing, video phone, video on demand are likely to be used extensively on internet and broadband networks. In fact, the word *multimedia* traffic mainly points to the presence of video traffic alongside data and voice traffic.

Today the VBR video codecs are mainly used. The variable-bit rate coding provides a constant quality and is supported by ATM. There are a number of compression methods for video, with MPEG (I and II) being the most used technique [12].

VBR video sources are highly bursty. The bit rate depends on the content of the scenes, the motions, and also the coding scheme. We may expect the trafic stream from a video conference to have less variation in bit rate than the movie *terminator*. Usually there is an abrupt change in bit rate when a scene change occures. Within a scene, only a small portion of the picture changes from a frame to the next frame. Also the nature of video data is such that it recorrelates at each frame and line interval. For lines, the reason is that the data on one part of an image line is very similar (or somehow correlated) to the data

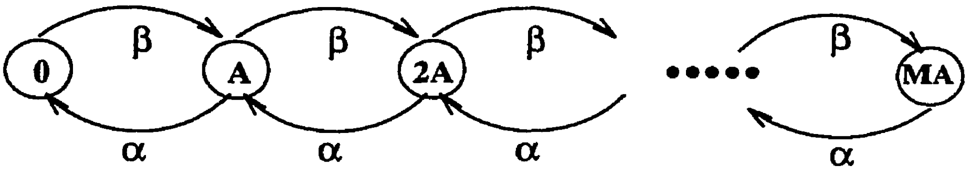

Figure 2.2: Discrete-state, continuous time Markov chain model for video

on the same part on the next line (which shows the same object). This property is called *spatial correlation.* For frames, within a scene the data in one part of the frame is highly correlated to the data in the same part in the next frame. It is called *temporal correlation.* Due to the above correlations, video traffic could not be modeled by a memoryless process such as Poisson.

There are several models for variable-bit rate video traffic [10]. Some models describe only the intrascene changes effectively. These models are more appropriate for videoconferencing, videophone or show talk programs where there are not many scene changes. To model the traffic of high motion movies one needs a model which captures scene changes too.

Here we briefly examine some of the various techniques for modeling of VBR video. For more details the reader may refer to [10] or [7].

### 2.6.1 Continuous time Discrete state Markov Model

This model was proposed by *Maglaris et al* in [14]. It is suitable for modeling the intrascene changes although it was later generalized to consider scene changes too.

The main idea here is to quantize the bit rate into finite discrete levels so that a continuous Markov chain as in Figure 2.2 can be formed. Then the chain can be broken down to a superposition of homogenous on-off minisources similar to the one in Figure 2.1 with $T = \frac{1}{A}$.

If the number of minisources is denoted by M, it is easy to show that for a superposition of M on-off minisources one can write [14] [1] :

$$\begin{gathered}
P[\lambda(t) = kA] = \frac{M!}{k!(M-k)!} p^k (1-p)^{M-k} \quad p = \frac{\beta}{\alpha+\beta} \\
E(\lambda) = MAp \\
C(0) = MA^2p(1-p) \\
C(\tau) = C(0)\, e^{-(\alpha+\beta)\tau}
\end{gathered} \tag{2.1}$$

Where $E(\lambda)$ and $C(\tau)$ are the average bit rate and autocovariance function of the superposed traffic, respectively. Therefore by measuring the average bit rate and autocovariance of the video traffic, the parameters $E(\lambda)$, $C(0)$ and $a = \alpha + \beta$ can be estimated. Then it is easy to show that:

$$\begin{gathered}
\alpha = a/[1 + \frac{E^2(\lambda)}{M\,C(0)}] \\
\beta = a - \alpha \\
A = \frac{C(0)}{E(\lambda)} + \frac{E(\lambda)}{M}
\end{gathered} \tag{2.2}$$

The only problem here is how to choose the number of minisources, M. In [14], the value of 20 has been proposed. However, in [10] it says a value of 8 also yields acceptable results. Surely the higher the number of minisources, the lower the quantization error.

A technique for queueing analysis of the above model is also presented in [14]. Therefore the model is analytically tractable.

### 2.6.2 Autoregressive models

Autoregressive models has been extensively used for modeling of video traffic [14]. In this class of continuous-state discrete-time models, the next random variable in the sequence is

[1] The notation here is slightly different from [14]. We changed the notation to keep it the same as Figure 2.1 and [17]. All of the equations have been changed respectively.

defined as an explicit function of the previous ones within a time window stretching from the present into the past [4]. The simplest case will be the linear Autoregressive model which is defined as follows:

$$\lambda(n) = a\lambda(n-1) + bw(n) \tag{2.3}$$

where $\lambda(n)$ represents the bit rate of the source during the $n$th frame, and $w(n)$ is a sequence of independent Gaussian noise. $a$ and $b$ are constants. Assuming that $w(n)$ has mean value of $\eta$ and variance of 1, the steady-state average and autocovariance function can be caluclated as [14]:

$$\begin{aligned} E(\lambda) &= \tfrac{b}{1-a}\eta \\ C(n) &= \tfrac{b^2}{1-a^2}a^n \quad n \geq 0 \end{aligned} \tag{2.4}$$

Therefore all of the parameters of the autoregressive model can be found by matching the average bit rate and autocorrelation function of the empirical traffic data to the above expressions.

Autoregressive models are suitable for modeling of intrascene changes in video traffic. The accuracy of the model will increase if the order of the model is increased. Although this class of models is suitable for simulation, but it is rather difficult to use them in queueing analysis. So their application is limited to traffic simulation.

A more complicated autoregressive model is ARMA (AutoRegressive Moving Average) model for video traffic, proposed in [15]. The advantage of ARMA model is in its ability to catch the *recorrelation* property, the one in which the autocorrelation curve has a number of peaks instead of a monotonic exponential decreasing. In ARMA model, the number of arrivals during $i$th interval, $X_i$, is given by:

$$X_i = g(\alpha Z_{i-m} + Y_i + v_i) \tag{2.5}$$

where $Z_i$ and $Y_i$ are sequences of correlated Gaussian noise random variables with mean zero, and $v_i$ is a sequence of uncorrelated Gaussian random variable. Check [15] for more information on the model parameter estimation.

Another class of models which fits into the group of autoregressive models is *Transform-Expand-Sample* (TES) model. The details of this model is beyond the scope of this thesis. For more information look at [4].

### 2.6.3 Markov-Modulated Poisson Process

The MMPP model is the main topic of research in this thesis and so we have dedicated an independent section to this model and two chapters to model parameter estimation. For more information on the MMPP model refer to Section 2.8. MMPP has been found to be a good model for representing a superposition of on-off sources ([17]) and therefore can be used as an alternative to discrete-time continuous-state Markov model which we discussed in Section 2.6.1. However, it has been shown that the MMPP is unable to catch some long range dependency effects in video traffic [10]. the MMPP model is short-range dependent, which means the effect of correlation is within short ranges. In other terms, the IDC curve of the MMPP model does not increase for long lags. In spite of this shortcoming, the analytical tractability of the MMPP queues makes them an attractive choice for modeling of various types of the traffic. We will describe the MMPP model in more detail in this thesis.

## 2.7 Data models

### 2.7.1 Properties of data traffic

Data traffic is the main type of *jitter-tolerant* traffic type. It means that unlike real time video and voice traffic, data traffic can tolerate a certain degree of delay variation, but it is very sensitive to cell loss, as it forces the sender to re-transmit. Therefore the most important QoS parameter for data traffic is the probability of loss, not average delay and delay variance.

Another feature of data traffic is that unlike video and voice, the statistical behaviour of data traffic is application-dependent. It means that it is impossible to come up with one universal data model and apply it to every case. Various types of data traffic such as WWW browsing, client-server transactions and LAN protocols each has different behaviour. It depends on the communication protocol, too. The performance of IP is expected to be different from X.25 or IPX. Due to the complexity and the large number of various situations and case studies for data traffic, the modeling of data traffic is still in its early ages. Nevertheless, there are certain characteristics which distinguish a data traffic stream. First of all, the data traffic is highly bursty, much burstier than video or voice. It is also long-range dependent. For more information on the definition of long-range and short-range dependencies refer to [7]. In our research we consider every traffic type with a forever-monotonically-increasing IDC curve (an infinite value for $IDC(\infty)$) to be long-range dependent. As it was explained before, the MMPP model is short-range dependent. In Figure 2.3 a comparison between IDC curves of short-range dependent and long-range dependent traffic is shown.

The new studies also reveal another property in data traffic, *self-similarity*, in LAN data traffic [8]. Self-similar processes display structural similarities across a wide range of

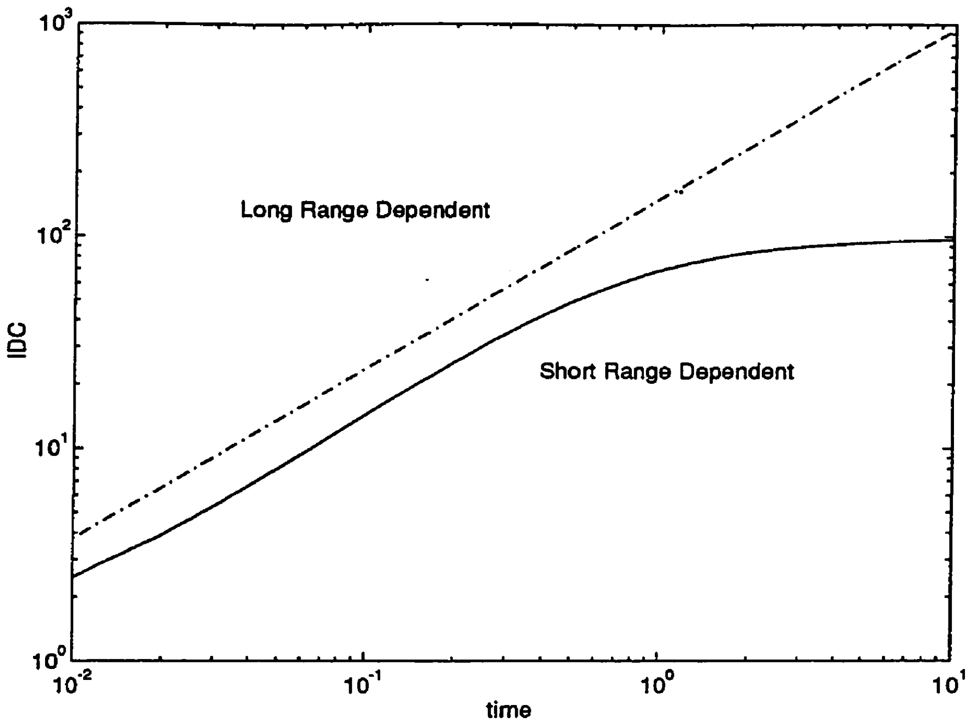

Figure 2.3: IDC curves of short range and long range dependent traffics

time scales. This property indicates the absence of a natural length of *burst*. At every time scale, ranging from a few miliseconds to minutes and hours, bursts consists of bursty subperiods separated by less bursty periods [7]. There are a number of models who show this property. Here we just very briefly describe one of them, *Pareto-Modulated Poisson Process*, or PMPP. For more information refer to [7].

### 2.7.2 PMPP Model

PMPP is an example of the models that show long-range dependency and so is believed to be able to represent ATM data traffic [6], [7]. The simplest version, 2-state PMPP, consists of a Poisson process switching between two average rates $\lambda_1$ and $\lambda_2$. The sojourn time in each state has a Pareto distribution, defined by the following probability density function:

$$f(t) = \alpha t^{-(\alpha+1)} \tag{2.6}$$

It has been shown that the IDC curve of this model will have a form of $1 + k\,t^{\beta}$, hence monotonically increasing with t [7].

PMPP is one of the models which are capable of catching the effect of long-range dependency. However, it is not analytically tractable yet. Therefore the use of the model is so far limited to simulations only.

## 2.8 Markov-Modulated Poisson Process (MMPP)

As we explained before, the main topic of this research is the applicability of the MMPP model in representing ATM traffic. Therefore here we study the model in more detail to provide the reader with some important characteristics of the MMPP model which we use in the rest of this thesis. The best source of information about the MMPP is [18].

The MMPP is a doubly-stochastic Poisson process whose arrival rate is given by an underlying m-state irreducible Markov chain which is independent of the arrival process. When Markov chain is in state $i$, arrivals occure according to a Poisson process of rate $\lambda_i$. The sojourn time at each state has an exponential distribution. In Figure 2.4 the states of the MMPP is shown for a simple 2-state MMPP.

The MMPP is parameterized by the Markov chain infinitesimal generator matrix $Q$ and

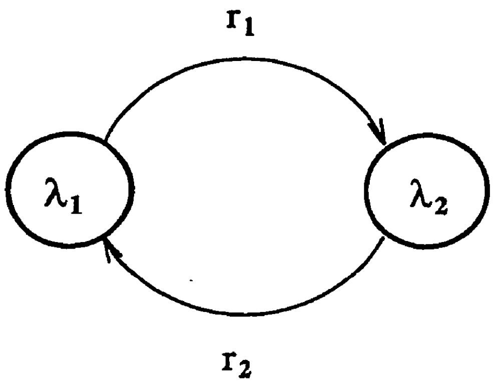

Figure 2.4: MMPP state diagram

Poisson arrival rate diagonal matrix Λ as follows:

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & -\sigma_2 & \cdots & \sigma_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \sigma_{m1} & \sigma_{m2} & \cdots & -\sigma_m \end{bmatrix} \tag{2.7}$$

$$\sigma_i = \sum_{j=1,\neq i}^{m} \sigma_{ij}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix} \tag{2.8}$$

$$\lambda = (\lambda_1,\ \lambda_2,\ \cdots,\ \lambda_m)^T \tag{2.9}$$

The steady-state vector of the Markov chain, $\pi$, can be computed from the following equation:

$$\pi Q = 0 \qquad \pi e = 0 \tag{2.10}$$

In the 2-state case, $\pi$ is given by

$$\pi = (\pi_1, \pi_2) = \frac{1}{r_1 + r_2}(r_2, r_1) \tag{2.11}$$

MMPP is not a renewal process, but it can be considered a Markov renewal process [18].

The superposition of MMPPs is again an MMPP. The generator $Q$ and rate matrix $\Lambda$ of the composite MMPP are calculated as follows:

$$\begin{aligned} Q &= Q_1 \oplus Q_2 \oplus \cdots \oplus Q_n \\ \Lambda &= \Lambda_1 \oplus \Lambda_2 \oplus \cdots \oplus \Lambda_n \end{aligned} \tag{2.12}$$

Where $\oplus$ represents the *Kronecker-sum* as defined in [18].

The IDC curve of the 2-state MMPP can be caluclated as follows [17]:

$$\mathrm{IDC}(t) = 1 + \frac{2r_1r_2(\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^2(\lambda_1 r_2 + \lambda_2 r_1)} - \frac{2r_1r_2(\lambda_1 - \lambda_2)^2}{(r_1 + r_2)^3(\lambda_1 r_2 + \lambda_2 r_1)t}(1 - e^{-(r_1+r_2)t}) \tag{2.13}$$

### 2.8.1 MMPP/G/1 queue

The Matrix Geometric techniques have been used for analysis of the MMPP/G/1 queue in [17] and [18]. Here we just summerize the algorithm very briefly. For more details please refer to [18].

*Inputs:*

- The transition rate matrix $Q$

- The cell generation rate matrix $\Lambda$

- Mean arrival rate $\lambda_{tot} = \pi\lambda$

- The service time distribution $\hat{H}(x)$ with finite mean $h$, second and third moments $h^{(2)}$ and $h^{(3)}$ and Laplace-Stieltjes transform $H(s)$

*Algorithm for solving MMPP/G/1 queue:*

1. Compute matrix $G$ as follows:

   - $G_0 = 0$, $H_{0,k} = I$, $k = 0, 1, 2, \cdots$,
   - $\Theta = max((\Lambda - Q)_{ii})$,
   - $\gamma_n = \int_0^\infty e^{-\Theta x} \frac{(\Theta x)^n}{n!} \mathrm{d}\hat{H}(x)$, $n = 0, 1, \cdots, n^*$

     where $n^*$ is chosen such that $\sum_{k=1}^{n^*} \gamma_k > 1 - \epsilon_1$, $\epsilon_1 \ll 1$.
   - For $k = 0, 1, 2, \cdots$ compute

$$H_{n+1,k} = [I + \tfrac{1}{\Theta}(Q - \Lambda + \Lambda G_k)]H_{n,k}, \qquad n = 0, 1, \cdots, n^*$$

$$G_{k+1} = \sum_{n=0}^{n^*} \gamma_n H_{n,k}$$

   - Continue the above recursion until $||G_{k-1} - G_k|| < \epsilon_2 \ll 1$
   - Set $G = G_{k+1}$

2. Compute the steady state vector $g$ which satisfies

$$gG = g \qquad ge = 1$$

3. Compute

$$x_0 = \frac{1-\rho}{\lambda_{tot}} g(\Lambda - Q)$$

4. Compute the system size distribution at departures

- Compute $A_v$ matrices as follows:

$$A_v = \sum_{n=v}^{\infty} \gamma_v K_v^{(n)} \qquad v \geq 0$$

$$\begin{aligned} K_0^{(0)} &= I, && \\ K_v^{(0)} &= 0, && v \geq 1 \\ K_0^{(n)} &= K_0^{(n-1)}[\Theta^{-1}(Q-\Lambda)+I], && v \geq 0 \\ K_v^{(n)} &= K_v^{(n-1)}[\Theta^{-1}(Q-\Lambda)+I] + K_{v-1}^{(n-1)}\Theta^{-1}\Lambda, && n \geq v \geq 1 \\ K_v^{(n)} &= 0, && n < 0, \ n < v \end{aligned} \tag{2.14}$$

The summation for $A_v$ must be truncated to the number $N$ which is chosen as a maximum of $N_1$ or $N_2$ where the following conditions are set for $N_1$ and $N_2$:

$$\sum_{n=0}^{N_1} \gamma_n \geq 1-\epsilon, \quad \max_j \left[ \sum_{v=0}^{N_2} (A_v e)_j - 1 \right] < \epsilon \tag{2.15}$$

$\gamma_n$ is defined in item 1.

-

$$\begin{aligned} \overline{A}_k &= A_k + \overline{A}_{k+1} G \\ B_k &= (\Lambda - Q)^{-1} \Lambda A_k \\ \overline{B}_k &= B_k + \overline{B}_{k+1} G \end{aligned} \tag{2.16}$$

It is a backward recursion. Therefore one must start at a sufficiently large index $i$ in order that $\sum_{k=i+1}^{\infty} A_k e$ and $\sum_{k=i+1}^{\infty} B_k e$ are negligible and so $A_i$ and $B_i$ could be set to zero.

- Compute system size distribution at departues as follows:

$$x_i = [x_0 \overline{B}_i + \sum_{v=1}^{i-1} x_v \overline{A}_{i+1-v}](1-\overline{A}_1)^{-1} \quad i \geq 1 \tag{2.17}$$

5. Compute

$$y_0 = (1-\rho)g$$

6. Compute the queue length distribution at an arbitrary time using the following equation:

$$y_i = [y_{i-1}\Lambda - \lambda_{tot}(x_{i-1} - x_i)](\Lambda - Q)^{-1} \tag{2.18}$$

7. The transform of waiting time distribution can be computed from the following equation:

$$W(s) = s(1-\rho)g[sI + Q - \Lambda(1 - H(s))]^{-1}e \tag{2.19}$$

In the next chapters we will present some simple formulas to approximate some part of the analysis of MMPP/D/1 queue as a special case.

Now we have enough information about MMPP model to start our discussion on MMPP parameter estimation, first for the simple two-state case and then for a more general multiple-state case in the next chapters.

# CHAPTER 3
# Modeling of Aggregate ATM Traffic using 2-state Markov Modulated Poisson Processes

## 3.1 Introduction

In this chapter we introduce some techniques for modeling various types of traffic by a 2-state Markov-Modulated Poisson Process (MMPP), introduced in Section 2.8. These techniques are widely used for deriving the corresponding model parameters for each type of ATM traffic, which we call *model parameter matching.*

We first start with one of the main parameters of MMPP model, the IDC (Index of dispersion for counts) curve. This parameter was introduced in the previous chapter. Here we look at it for special case of MMPP model in more detail. IDC curve plays a major role in queueing performance and therefore is extensively used for parameter matching purposes. Then we start the MMPP parameter matching process for a special type of ATM traffic, the simple case of aggregated voice traffic. We study and compare several different techniques proposed for parameter matching in this case. Then we extend it to a general case of ATM traffic, known only by its samples. We present a new, refined matching technique for modeling of an arbitrary ATM traffic by a 2-state MMPP model.

### 3.2 The IDC curve for 2-state MMPP model

In the previous chapter we defined the index of dispersion for counts as the variance of the number of arrivals over an observation interval, divided by the mean of the number of arrivals over the same fixed observation interval. It has been analytically shown that for a 2-state MMPP model with parameters [ $\lambda_1$ , $\lambda_2$ , $r_1$ , $r_2$ ], The IDC curve is derived as follows [17]:

$$\mathrm{IDC}(t) = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2 (\lambda_1 r_2 + \lambda_2 r_1)} - \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^3 (\lambda_1 r_2 + \lambda_2 r_1)\, t}(1 - e^{-(r_1+r_2)t}) \quad (3.1)$$

Equation (3.1) can be re-written in the following simpler form:

$$\mathrm{IDC}(t) = \mathrm{IDC}(\infty) - \frac{\mathrm{IDC}(\infty) - 1}{d\,t}(1 - e^{-d\,t}) \quad (3.2)$$

where:

$$\mathrm{IDC}(\infty) = 1 + \frac{2(\lambda_1-\lambda_2)^2 r_1 r_2}{(r_1+r_2)^2 (\lambda_1 r_2+\lambda_2 r_1)} \qquad d = r_1 + r_2$$

Equation (3.2) shows that the IDC curve of the 2-state MMPP source has only two parameters, IDC($\infty$) and $d = r_1 + r_2$. Figures 3.1 and 3.2 show the effect of each of the parameters on the resulted IDC curve.

The parameters of IDC curve have very important effects on the queueing performance. In order to examine this effect, we generate some MMPP model with particular IDC curves and compare their performance in an MMPP/D/1 queue. The measure of performance is the probability of loss, which we approximate here with the survivor function of the queue length for an infinite buffer. In all of the following examples, the mean arrival rate $\overline{\lambda}$ is kept the same. The traffic load is 90% for all of the MMPP/D/1 queues.

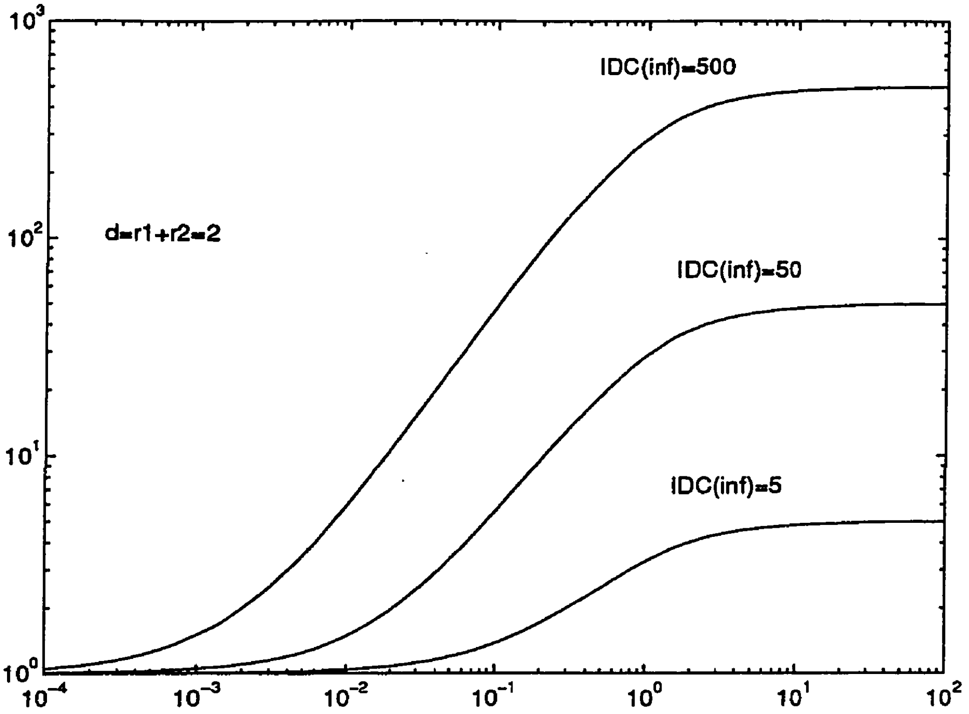

Figure 3.1: IDC curves with different values of IDC($\infty$)

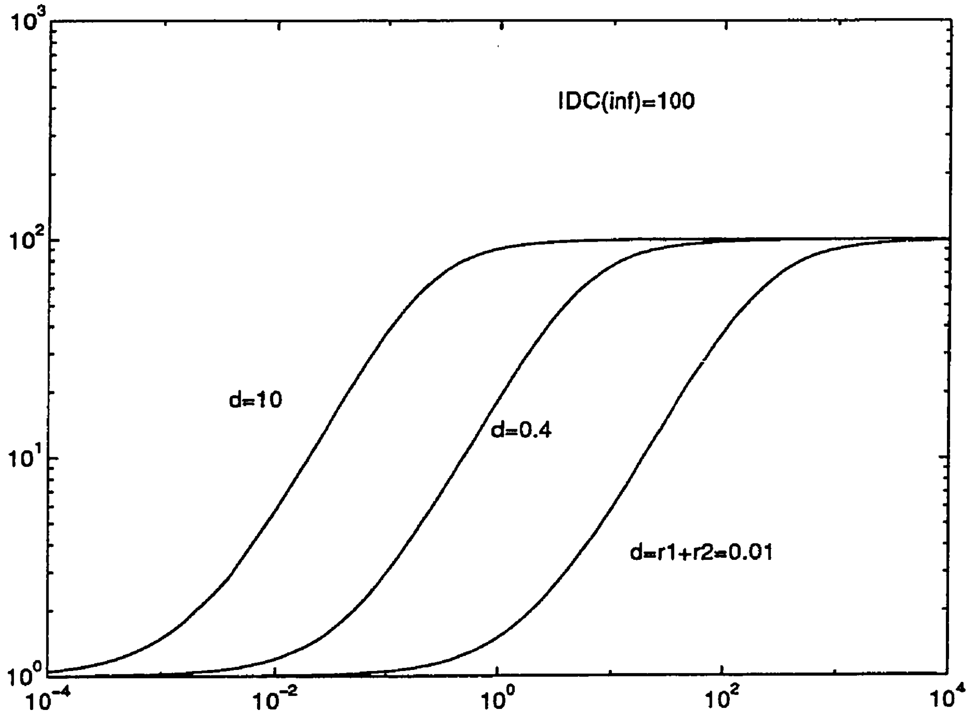

Figure 3.2: IDC curves with different values of $d = r_1 + r_2$

### 3.2.1 The effect of IDC($\infty$)

Let us suppose that we have two models with exactly the same mean arrival rate and $d$ but with different IDC($\infty$). Assume that the new model has a value of IDC($\infty$) of K times of that of the reference model. The parameters of the new model $[\hat{\lambda}_1 \ , \ \hat{\lambda}_2 \ , \ \hat{r}_1 \ , \ \hat{r}_2]$ can be calculated from the original parameter set by using the following equations:

$$\begin{array}{cc} \hat{r}_1 = r_1 & \hat{r}_2 = r_2 \\ \hat{\lambda}_1 = \frac{(\sqrt{K} r_1 + r_2)\lambda_1 + r_1(1-\sqrt{K})\lambda_2}{r_1 + r_2} & \hat{\lambda}_2 = \hat{\lambda}_1 - \sqrt{K}(\lambda_1 - \lambda_2) \end{array} \tag{3.3}$$

The following table shows a sample case:

| Process | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ ($s^{-1}$) | $r_2$ ($s^{-1}$) |
|---|---|---|---|---|
| P1 | 6651.7 | 5359.6 | 1.7333 | 1.0783 |
| P2 | 6107.0 | 5698.4 | 1.7333 | 1.0783 |
| P3 | 8374.1 | 4288.1 | 1.7333 | 1.0783 |

Table 3.1: Three MMPP models with different IDC($\infty$)

In Figure 3.3 the performance of the models in an MMPP/D/1 queue are compared. A huge difference is noticed. It shows that the larger the value of IDC($\infty$), the higher the probability of cell loss.

### 3.2.2 The effect of $d = r_1 + r_2$

Here we have two models the same mean arrival rate $\overline{\lambda}$ and IDC($\infty$) but with different $d$. Assume that the new model has a value of IDC($\infty$) as K times of the origical model. It

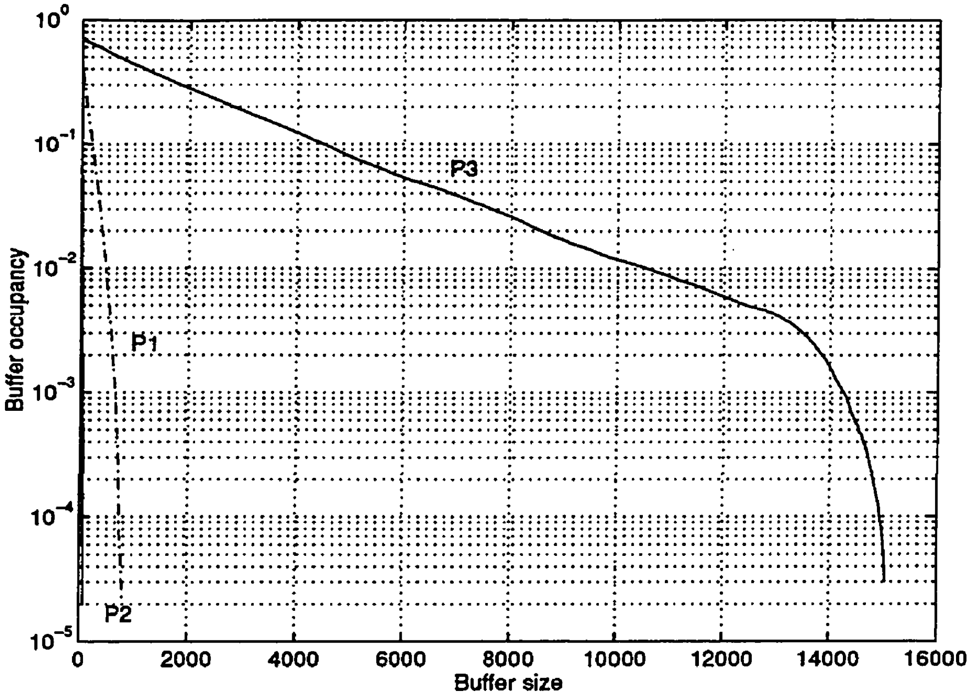

Figure 3.3: The effect of IDC($\infty$) on the queueing performance

is easy to show that in this case the values of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ can be derived from Equation set (3.3). For transition rates we can write:

$$\hat{r}_1 = K\,r_1 \quad \hat{r}_2 = K\,r_2 \tag{3.4}$$

The test case model parameters are shown in Table 3.2.2. The queueing performance of the models is shown in Figure 3.4. Again here, the increase in $d$ results in an increase in the probability of cell loss.

### 3.2.3 Models with the same IDC curve

It is also possible to generate two MMPP model with the same IDC curve but with different queueing performance. The trick is to keep $\overline{\lambda}$, IDC($\infty$) and $d$ the same for both of the models but to change the value of $r_1/r_2$. Here suppose that we increase the value of $r_1/r_2$ by a

| Process | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ ($s^{-1}$) | $r_2$ ($s^{-1}$) |
|---|---|---|---|---|
| P1 | 6651.7 | 5359.6 | 1.7333 | 1.0783 |
| P2 | 13821.0 | 899.705 | 173.3264 | 107.8307 |
| P3 | 6107.0 | 5698.4 | 0.1733 | 0.1078 |
| P4 | 8374.1 | 4288.1 | 17.3326 | 10.783 |

Table 3.2: Four MMPP models with different values of $d = r_1 + r_2$ but the same IDC($\infty$) and mean arrival rate

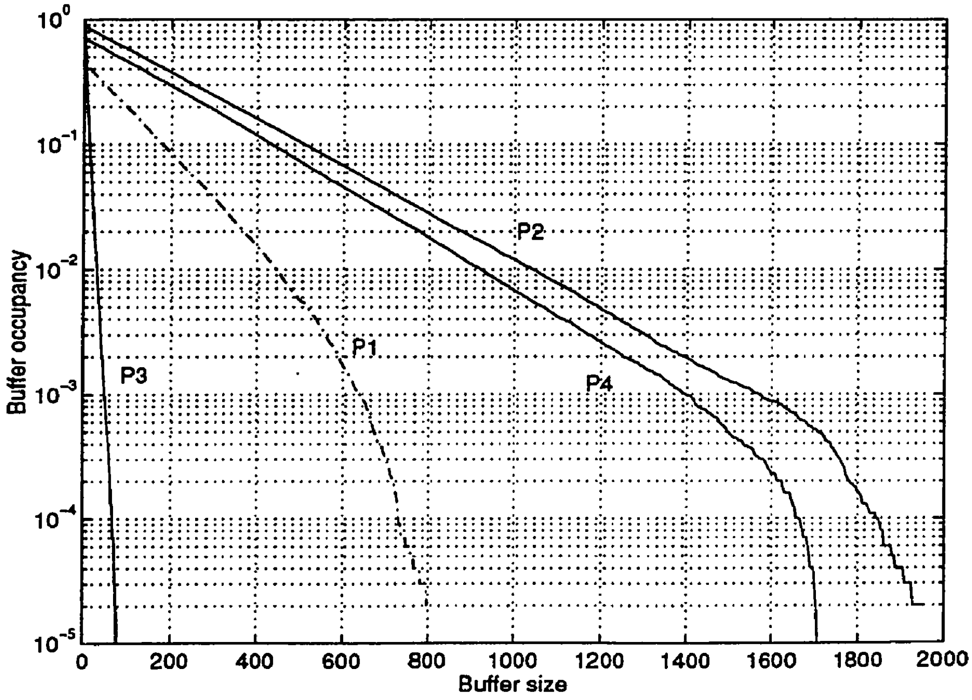

Figure 3.4: The effect of $d = r_1 + r_2$ on the queueing performance

factor of K. Then the new model parameters can be calculated from the original parameter set as follows:

$$\begin{aligned}
&\hat{r}_1 = \tfrac{K(r_1+r_2)}{K+1} \qquad \hat{r}_2 = \tfrac{r_1+r_2}{K+1} \\
&\overline{\lambda} = \tfrac{\lambda_1 r_2+\lambda_2 r_1}{r_1+r_2} \\
&D = \tfrac{\lambda_1-\lambda_2}{r_1+r_2}\sqrt{\tfrac{r_1 r_2}{K}} \\
&\hat{\lambda}_1 = \overline{\lambda} + K\,D \qquad \hat{\lambda}_2 = \overline{\lambda} - D
\end{aligned} \tag{3.5}$$

In Table 3.2.3 a test case is shown with some 2-state MMPP models with exactly the same IDC curve and mean. The result of MMPP/D/1 simulation is shown in Figure 3.5 which indicates a big difference in the performance.

| Process | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ ($s^{-1}$) | $r_2$ ($s^{-1}$) |
|---|---|---|---|---|
| P1 | 6651.7 | 5359.6 | 1.7333 | 1.0783 |
| P2 | 7841.9 | 5656.5 | 2.556 | 0.2556 |
| P3 | 6053.8 | 3868.4 | 0.2556 | 2.556 |

Table 3.3: Three MMPP models with exactly the same IDC curves

The bottom line of the above results is that, while IDC curve plays a very important role in the queueing performance, it is not enough for the unique identification of the model parameters. As we are going to show in the next sections, some other parameters of the traffic must be used too.

Now let us start our model parameter matching. First we study the simplest case, the modeling of aggregated voice sources. We call it the simplest case because a very well studied mathematical model for PCM voice source is available: On-Off source.

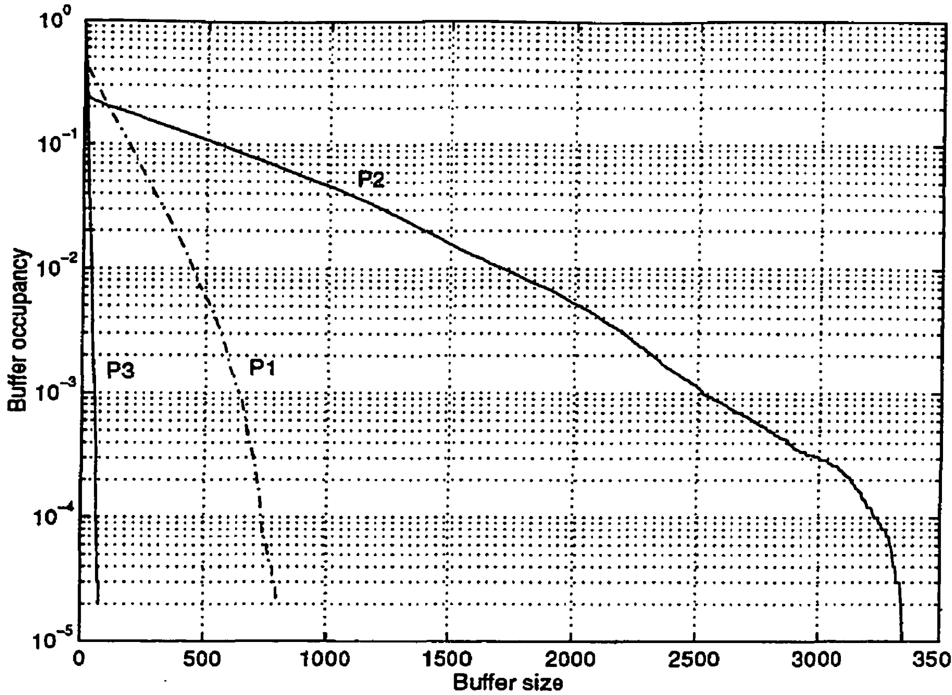

Figure 3.5: Models with the same mean and IDC curves but different queueing performance

### 3.3 Modeling of aggregated voice traffic by a 2-state MMPP

In Section 2.5.1 the on-off model for voice source and its various parameters were described. Figure 2.1 shows the model. In this section we examine the case of aggregated voice sources which can be modeled by a superposition of on-off sources.

In most of the cases, a (large) number of voice sources are multiplexed on the same line before reaching the ATM switch. So the problem is how to solve a queue or switch with a superposition of on-off sources as input. One proposed alternative is to match a two-state MMPP model as appears in Figure 2.4 to the superposition of on-off sources.

Several techniques have been proposed for deriving the parameters of the MMPP model to be matched to the aggregated on-off sources. A range of the charcateristics of MMPP and on-off sources are used in the matching, such as moments of arrival rates or interarrival times. Here we go over some of the more famous techniques briefly. We use the following assumptions:

- The voice sources are packetized.
- All of the voice sources have identical parameters, or in other words, are *homogeneous.* Although a few techniques work in hetrogeneous case too.
- Only *active* sources are considered, in other words, those sources who currently are holding a call. Usually the process of making a call is modeled by a Poisson process. Call admission process is not a topic of interest in this research.

The following notation for the parameters for the superposed on-off traffic and MMPP model is used:

- Parameters of the superposition of on-off sources:
  - $N$ : Number of active on-off sources
  - $\alpha$ : Mean transition rate from state ON to state OFF
  - $\beta$ : Mean transition rate from state OFF to state ON
  - $T$ : Fixed interarrival time in state ON
  - $A$ : Fixed cell generation rate in state ON, equals to $1/T$
- Parameters of 2-state MMPP model:
  - $\lambda_1$ : Mean cell generation rate in state 1
  - $\lambda_2$ : Mean cell generation rate in state 2
  - $r_1$ : Mean transition rate from state 1 to state 2
  - $r_2$ : mean transition rate from state 2 to state 1

The models have been shown in Figures 2.1 and 2.4.

### 3.3.1 Moment-based matching

The moment-based matching was first introduced by Heffes and Lucantoni in [17]. The technique has been widely used thereafter and has been extended to more general cases as well. In this technique, the four parameters of the 2-state MMPP are chosen so that the following characteristics of the superposition of on-off sources are matched with those of 2-state MMPP model:

1. The mean arrival rate
2. The IDC (variance-to-mean ratio of the number of arrivals in the interval $(0, t_1)$ )
3. The asymptotic value of IDC (IDC($\infty$))
4. The third central moment of the number of arrivals in the interval $(0, t_2)$

Now let us calculate the value of each of the above characteristics for superposition of on-off sources and 2-state MMPP in term of the model parameters. We just briefly offer the results of the matching technique. For detailed mathematical derivation check [17].

For a single on-off source, the moments of number of arrivals over an interval which we denote by random variable $N(0:t)$, can be defined as:

$$M_r(t) = E[N^r(0:t)] \tag{3.6}$$

The Laplace transform of the moments, $\hat{M}_r(s)$ can be calculated for the first three moments as follows:

$$\begin{aligned} \hat{M}_1(s) &= \lambda/s^2 \\ \hat{M}_2(s) &= \tfrac{\lambda}{s^2}\left(\tfrac{1+\hat{f}(s)}{1-hatf(s)}\right) \\ \hat{M}_3(s) &= \tfrac{\lambda}{s^2}\left(\tfrac{1+4\hat{f}(s)+\hat{f}^2(s)}{(1-\hat{f}(s))^2}\right) \end{aligned} \tag{3.7}$$

where $\lambda = 1/(T+\alpha T/\beta)$ is the mean arrival rate and $\hat{f}(s) = [1-\alpha T+\alpha T\,\beta/(s+\beta)]\,e^{-sT}$ is the Laplace-Stjelties Transform of the interarrival distribution. Obviously, $M_1(t) = \lambda\,t$.

The IDC curve is defined as $\text{var}[N(0:t)]/E[N(0:t)]$ and can be derived using $M_1$ and $M_2$. In [17] it was shown that the value of IDC at large lags for the superposition of on-off sources could be calculated as:

$$\lim_{t\to\infty} \text{IDC}(t) = \text{IDC}(\infty) = \frac{1-(1-\alpha T)^2}{(\alpha T+\beta T)^2} \tag{3.8}$$

Also the third central moment for the superposition process can be calculated from $M_i$s defined in (3.7) as follows:

$$\mu_3^S(0,t) = n\,[M_3(t) - 3M_2(t)M_1(t) + 2M_1^3(t)] \tag{3.9}$$

For the 2-state MMPP, If we denote $N_t$ as the number of arrivals of the stationary 2-state MMPP over the interval $(0,t)$, The moments of $N_t$ can be calculated as follow:

$$\begin{aligned} \overline{N}_t &= E[N_t] = \tfrac{\lambda_1 r_2+\lambda_2 r_1}{r_1+r_2}\,t \\ \text{IDC}(t) &= \tfrac{\text{var}(N_t)}{E[N_t]} = 1 + \tfrac{2\,(\lambda_1-\lambda_2)^2\,r_1\,r_2}{(r_1+r_2)^2\,(\lambda_1\,r_2+\lambda_2\,r_1)} - \tfrac{2\,(\lambda_1-\lambda_2)^2\,r_1\,r_2}{(r_1+r_2)^3\,(\lambda_1\,r_2+\lambda_2\,r_1)\,t}(1-e^{-(r_1+r_2)t}) \\ \text{IDC}(\infty) &= 1 + \tfrac{2\,(\lambda_1-\lambda_2)^2\,r_1\,r_2}{(r_1+r_2)^2\,(\lambda_1\,r_2+\lambda_2\,r_1)}. \end{aligned} \tag{3.10}$$

$$E[(N_t - \overline{N}_t)^3] = g^{(3)}(1,t) - 3\overline{N}_t(\overline{N}_t - 1)\frac{\mathrm{var}(N_t)}{\overline{N}_t} - \overline{N}_t(\overline{N}_t - 1)(\overline{N}_t - 2) \tag{3.11}$$

where

$$\begin{aligned}
g^{(3)}(1,t) &= \tfrac{6}{r_1+r_2}\left[\tfrac{A_{11}}{6}t^3 + \tfrac{A_{21}}{2}t^2 + A_{31}t + A_{12}t\,e^{-(r_1+r_2)t} + A_{41}(1 - e^{-(r_1+r_2)t})\right] \\
A_{11} &= \tfrac{(\lambda_1 r_2+\lambda_2 r_1)^3}{(r_1+r_2)^2} \\
A_{21} &= \tfrac{2r_1r_2(\lambda_1-\lambda_2)^2(\lambda_1 r_2+\lambda_2 r_1)}{(r_1+r_2)^3} \\
A_{31} &= \tfrac{r_1r_2(\lambda_1-\lambda_2)^2[\lambda_1r_1+\lambda_2r_2-2(\lambda_1r_2+\lambda_2r_1)]}{(r_1+r_2)^4} \\
A_{41} &= \tfrac{-2r_1r_2(\lambda_1-\lambda_2)^3(r_1-r_2)}{(r_1+r_2)^5} \\
A_{12} &= \tfrac{r_1r_2(\lambda_1-\lambda_2)^2(\lambda_1r_1+\lambda_2r_2)}{(r_1+r_2)^4}
\end{aligned} \tag{3.12}$$

In the following the algorithm for finding the parameters of the 2-state MMPP model from the parameters of the superposition of On-Off sources is explained:

1. Parameters of Aggregate on-off sources: $\alpha$, $\beta$, $T$, $N$.

   Parameters of the 2-state MMPP model: $\lambda_1$, $\lambda_2$, $r_1$, $r_2$.

2. Considering Equations (3.7), we define:

$$\begin{aligned}
a &= \tfrac{M_1(t)}{t} = \lambda = 1/(T + \alpha T/\beta) \\
b_t &= \mathrm{IDC}(t) = \tfrac{M_2(t)-M_1^2(t)}{M_1(t)} \\
b_\infty &= \mathrm{IDC}(\infty) = \tfrac{1-(1-\alpha T)^2}{(\alpha T+\beta T)^2}
\end{aligned} \tag{3.13}$$

3. Calculate $b_t$ at an arbitrary chosen point $t_1$ and find $d = r_1 + r_2$ from the following nonlinear equation numerically:

$$d = \frac{1}{t_1}\left(\frac{b_\infty - 1}{b_\infty - b_{t_1}}(1 - e^{-dt_1})\right) \tag{3.14}$$

4. Using (3.9), calculate third central moment $\mu_3^S(0:t)$ at an arbitrary time lag $t_2$. Define $K = (\lambda_1 - \lambda_2)(r_1 - r_2)$, and find K from the following equation:

$$\begin{gathered}\mu_3^S(0:t) + 3at_2(at_2-1)b_{t_2} + at_2(at_2-1)(at_2-2) = \\ a^3t_2^3 + 3a^2(b_\infty - 1)t_2^2 + \tfrac{3a(b_\infty-1)}{d}\left[\tfrac{K}{d} - a\right]t_2 + \tfrac{3a}{d^2}(b_\infty - 1)(K+ad)t_2e^{-dt_2} \\ -\tfrac{6a}{d^3}(b_\infty - 1)K(1-e^{-dt_2})\end{gathered} \tag{3.15}$$

5. Then based on the value of K, we will have:

- If K=0,

$$\begin{gathered}r_1 = r_2 = \tfrac{d}{2} \\ \lambda_1 = a + \tfrac{1}{2}\sqrt{2ad(b_\infty - 1)} \\ \lambda_2 = a - \tfrac{1}{2}\sqrt{2ad(b_\infty - 1)}\end{gathered} \tag{3.16}$$

- If $K \neq 0$ then we define $e = \frac{(b_infty-1)ad^3}{2K^2}$ and write:

$$\begin{gathered}r_1 = \tfrac{d}{2}\left(1 + \tfrac{1}{\sqrt{4e+1}}\right) \\ r_2 = d - r_1 \\ \lambda_2 = \left(\tfrac{ad}{r_2} - \tfrac{K}{r_1-r_2}\right)\left(\tfrac{r_2}{r_1+r_2}\right) \\ \lambda_1 = \tfrac{K}{r_1-r_2} + \lambda_2\end{gathered} \tag{3.17}$$

The time lags $t_1$ and $t_2$ may be chosen arbitrarily. However, it is better to choose them in a way that we get a good fit of IDC curve.

In [17] a technique for solving MMPP/G/1 queue has been proposed too which we reviewed in Section 2.8.1. The performance of the moment-based technique against other techniques will be studied in Section 3.3.4.

The moment-based matching, as we are going to show later, offers a very good matching between 2-state MMPP model and the superposition of on-off sources. It matches IDC

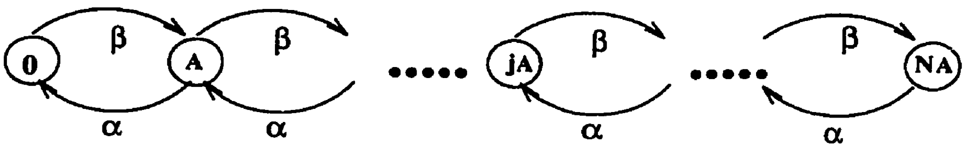

Figure 3.6: Birth-death process representing the superposition of on-off sources

curve very well. The main problem with the moment-based matching technique lays in the difficult, lengthy calculations of inverse Laplace transform.

### 3.3.2 Overload-Underload Approach

The idea of underload-overload approach has been used in the matching techniques in [23], [29] and [22] among others.

The approach is simple. Let us consider a superposition of N independant and homogenous on-off sources as we described before. Such superposition results in a birth-death process whose states show how many of the sources are in ON state. If we denote the state by $J(t)$, the Markov chain of the process has $N+1$ states, from $J(t)=0$ up to $J(t)=N$. The probability of being at state j in steady state, is equal to probably of having j out of N sources in ON state, clearly a binomial distribution. For single On-off source, the steady-state probability of being in ON state is equal to $\frac{\beta}{\alpha+\beta}$ where $\alpha$ is the mean transition rate from ON state to OFF state ($\alpha^{-1}$ is the mean sojourn time in ON state) and $\beta$ denotes the mean transition rate from OFF state to ON state ($\beta^{-1}$ is the mean sojourn time in OFF state). Then for the Markov chain representing the superposition of on-off sources, If we denote the steady-state probability of staying at state j by $\pi_j$, we can write:

$$\pi_j = \frac{N!}{j!\,(N-j)!}\left(\frac{\beta}{\alpha+\beta}\right)^j\left(\frac{\alpha}{\alpha+\beta}\right)^{(N-j)} \tag{3.18}$$

Figure 3.6 shows the corresponding Markov chain.

Now if we want to model the above chain with a two-state MMPP, simple physical consideration suggests that we can divide the states of the phase process into two subsets: an *overload* region, and an *underload* region so that each of the states of the approximating two-state MMPP corresponds to one of the regions. The border of the regions (or the threshold of the overload state) can be decided in various ways. One suggestion is to use the *mean number of sources in ON state* as the threshold. In this case, assuming that the overload region starts from state M+1, we can write:

$$M = \lfloor \frac{N\beta}{\alpha + \beta} \rfloor \tag{3.19}$$

where N denotes the number of active sources. Therefore, the *underload* region comprises the states {0, 1, ..., M} and the *overload* region comprises the states {M+1, ..., N}.

Now in order to find the parameters of the two-state MMPP model from the parameters of the superposition of on-off sources, we require the mean arrival rate at state 1 of the MMPP, $\lambda_1$, to be equal to the mean arrival rate in the overload region of the phase process. Similarly, $\lambda_2$ is equal to the mean arrival rate in the underload region of the phase process. Considering that the arrival rate at each state of the process is fixed (as Figure 3.6 shows it), we can write:

$$\lambda_1 = \sum_{i=M+1}^{N} i\,A\,\frac{\pi_i}{\pi_{OL}} \qquad \lambda_2 = \sum_{i=0}^{M} i\,A\,\frac{\pi_i}{\pi_{UL}} \tag{3.20}$$

where:

$$\pi_{OL} = \sum_{i=M+1}^{N} \pi_i \qquad \pi_{UL} = \sum_{i=0}^{M} \pi_i \tag{3.21}$$

where $\pi_i$ is defined by Equation (3.18).

The above equations determine the mean arrival rates at each of the two states of the

MMPP model. In order to calculate the mean transition rates, several different approaches have been taken by different researchers. Here we present three techniques. In all of them, the values of mean arrival rate are calculated by Equation (3.20).

- *Asymptotic matching*

  This technique was proposed in [23]. If we denote by random variable $\tau$ an overload period duration in the phase process, the survivor function of $\tau$ will have an exponential form like $G_\tau(x) = p_0 \exp(Q\,x)e$ for $x \geq 0$, in which $Q$ is the $(N-M) \times (N-M)$ transition rate matrix for overload region, and $p0 = [1,\ 0,\ \ldots,\ 0]$ denotes the initial probability distribution of the transient states. It could be shown that there exists one dominant eigenvalue of $Q$ which is real and negative. Therefore if we denote it by $\eta$, we will have:

  $$G_\tau(x) = D\,e^{-\eta x} + O(e^{-\eta x}) \tag{3.22}$$

  Using this approximation, we choose $r_{OL} = r_1 = \eta$. Therefore $r_1$ can be calculated from the maximal real-part eigenvalue of $Q$. Now by equating the mean arrival rate for MMPP and aggregated voice sources, $r_2$ can be easily calculated as:

  $$r_2 = r_1 \frac{\overline{\lambda} - \lambda_2}{\lambda_1 - \overline{\lambda}} \tag{3.23}$$

  where $\lambda_1$ and $\lambda_2$ are calculated from Equation (3.20) and $\overline{\lambda} = N\,A\,\beta/(\alpha + \beta)$ is the mean arrival rate for the aggregated voice sources.

- $\sum$*-matching*

  Introduced in [29], this technique is very similar to asymptotic matching and is primarily used for heterogeneous case where the parameters of the on-off sources are not identical but fit into several *classes.* However the technique can be simplified for homogeneous case. Like asymptotic matching, here also the mean sojourn time of the phase process in overload region is equated to mean sojourn time in state 1 of the MMPP, and that of the underload region to the mean sojourn time in state 2. A recursive formula is employed for determining the transition rates as follows [1]:

$$U_{k+1} = \begin{cases} \frac{1}{(M+1)\alpha} & ,k = M \\ \frac{(k-M)(1+\beta/\alpha)}{k+1} U_k + \frac{1}{(k+1)\beta} & ,M+1 \leq k \leq N-1 \end{cases} \tag{3.24}$$

$$r_{OL} = r_1 = \frac{1}{U_N}$$

$$r_{UL} = r_2 = \frac{\pi_{OL}}{\pi_{UL}} r_1$$

where $\pi_{OL}$ and $\pi_{UL}$ can be calculated from Equation (3.21). the mean arrival rates may be determined from (3.20). In Section 3.3.4 the performance of $\sum$-matching technique is compared to some other techniques.

$\sum$-matching technique has the advantage of the applicability in the heterogeneous case. Furthermore, it uses very simple calculations for mean arrival rates based on overload-underload assumption and a relatively simple, iterative formula for mean transition rates. However, this technique, as we will show later, does not match IDC curve. The MMPP model based on this technique will fail to predict the queueing performance under heavy traffic load, consequently.

- *IDC matching*

In this technique [22], instead of the time process, the counting process is considered, so the random variable to be used here is the number of arrivals over a fixed observation

[1] In order to keep consistency in this thesis, the definition of the parameters of on-off source $\alpha$ and $\beta$ have been changed from the original text ([29]), so are all of the equations.

interval. The mean arrival rates are calculated from (3.20). Then for mean transition rates we will need two more equations. We equate the mean arrival rates for both MMPP and agrregated on-off models, and the value of IDC($\infty$) for both models. Therefore we have the following set of equations:

$$\begin{aligned} \frac{\lambda_1 r_2+\lambda_2 r_1}{r_1+r_2} &= \frac{\beta}{\alpha+\beta} N\,A \\ \frac{2(\lambda_1-\lambda_2)^2\, r_1\, r_2}{(r_1+r_2)^2\,(\lambda_1\, r_2+\lambda_2\, r_1)} &= \frac{1-(1-\alpha T)^2}{(\alpha T+\beta T)^2} \end{aligned} \tag{3.25}$$

After solving the above set of equations for $r_1$ and $r_2$, the mean transition rates are calculated as follow:

$$\begin{aligned} r_1 = r_{OL} &= \frac{2(\lambda_{OL}-\lambda_{avg})^2(\lambda_{avg}-\lambda_{UL})}{(\lambda_{OL}-\lambda_{UL})\,\lambda_{avg}\,(IDC(\infty)-1)} \\ r_2 = r_{UL} &= \frac{2(\lambda_{OL}-\lambda_{avg})(\lambda_{avg}-\lambda_{UL})^2}{(\lambda_{OL}-\lambda_{UL})\,\lambda_{avg}\,(IDC(\infty)-1)} \end{aligned} \tag{3.26}$$

where $\lambda_{avg}$ denotes the mean arrival rate for aggregated on-off sources and $\lambda_{OL} = \lambda_1$ and $\lambda_{UL} = \lambda_2$ are calculated from (3.20).

The IDC matching technique enjoys the advantage of simplicity even more than $\sum$-matching, because the procedure to calculated the mean transition rates is simpler. Furthermore, the model captures the effect of the correlation in the traffic too, so in some sense it combines the advantages of each of the previos techniques and avoids their drawbacks. In Section 3.3.4 we will compare its performance against other techniques and show its capabilities.

### 3.3.3 Other matching techniques

In this section we are going to review some other techniques for matching of a 2-state MMPP model to a superposition of on-off sources very briefly.

Most of the techniques simply replace one of the matched qualities. It was mathematically proved that it is impossible to characterize a 2-state MMPP model only by the first two moments of its counting or time process. Moment-based matching ([17]) uses three moments. However, it has been also proved that with a combination of the first and second moments of the counting and the time process, the 2-state MMPP model is characterizable [25]. This fact has led many researchers to look into various combinations for matching purposes.

In [24] a purely interarrival time-based matching technique has been proposed. The technique matches the autocovariance function and the complementary probability distribution function of the interarrival times for both 2-state MMPP and superposition of voice sources. The technique is measurement based, means that some parameters must be measured for the aggregated voice traffic so that a 2-state MMPP could be matched to it. The following assumptions are made:

- The complementary probability distribution function of the interarrival time ($\Pr(X_i > x)$) for 2-state MMPP model has a 2nd-order hyperexponential format like $F_c(x) = q\, e^{-u_1 x} + (1-q)\, e^{-u_2 x}$.
- The autocovariance function of the interarrival time $C[k]$ has an exponential format like $C[k] = A\, \sigma^k$.

So therefore by measuring the parameters $u_1$, $u_2$, $q$ and $\sigma$ for the aggregated traffic, one can calculate the parameters of 2-state MMPP model as follows [24]:

$$\begin{gathered}
\lambda_1 = \tfrac{1}{2}\left[q(1-\sigma)(u_1-u_2)+\sigma u_1+u_2+\sqrt{[q(1-\sigma)(u_1-u_2)+\sigma u_1+u_2]^2-4\sigma u_1 u_2}\right] \\
\lambda_2 = \tfrac{u_1 u_2[\lambda_1-q(u_1-u_2)-u_2]}{\lambda_1 u_1-\lambda_1 q(u_1-u_2)-u_1 u_2} \\
r_1 = \tfrac{(u_1-\lambda_1)(u_2-\lambda_1)}{\lambda_2-\lambda_1} \\
r_2 = \tfrac{(\lambda_2-u_1)(\lambda_1+r_1-u_1)}{u_1-\lambda_1}
\end{gathered} \tag{3.27}$$

The fact that this technique uses measurement for matching, may indicate that one can use it for a general arbitrary case too (refer to the next section), however, as far as the issue of modeling the aggregate voice traffic is concerned, this fact will be a major drawback because the matching process cannot be done without samples from the traffic. Furthermore, the assumptions which it makes is in general not valid for a superposition of on-off sources. On-off sources do not have exponential autocovariance function. In [24], the IPP source has been used in place of on-off source which as we showed is in fact a special case of 2-state MMPP and is not used for modeling of fixed-rate PCM voice source.

Another technique was proposed in [25] which we will describe in Section 3.4.2.1.

### 3.3.4 Comparison of the performance of the matching techniques

In this section we compare the performance of some of the techniques for the matching of a 2-state MMPP to a superposition of on-off sources. We picked three of the techniques, *moment-based technique*, *IDC matching technique* and $\sum$*-matching technique.* We use simulations for our study. In order to compare the matching performance, we build two cases as follow:

1. In the first case, we form a G/D/1 queue in which the input consists of a superposition of on-off sources with known parameters, namely, $\alpha$, $\beta$, $A$ and $N$ number of the sources.
2. In our second case, we apply each of the above-mentioned matching techniques to derive the parameters of a 2-state MMPP model, and then we simulate the performance of the 2-state MMPP/D/1 queue in the same way as Case 1.

For Case 1, the parameters of the aggregated on-off sources are listed in the following table:

| No. of Sources | $\alpha$ | $\beta$ | A |
|---|---|---|---|
| 100 | 2.8409 ($s^{-1}$) | 1.5385 ($s^{-1}$) | 166.67 (Cells/S) |

Table 3.4: The parameters of the aggregated voice sources

The values of the mean transition rates have been picked from [17] and [30] and indicate the mean sojourn time of 0.352 ms in ON state and 0.650 ms in OFF state. The value of A (the fixed cell generation rate in ON state) corresponds to a 64kbps PCM voice line, assuming 48-byte long cells.

By applying each of the matching techniques, we get a different set of parameters for our 2-state MMPP model. The corresponding model parameters for each of the matching techniques are listed in Table 3.3.4. For the moment-based matching technique, both of the IDC curve and the third central moment have been matched to those of the traffic from the superposed on-off sources at t=0.5 s.

As Table 3.3.4 shows, both $\sum$-matching and IDC matching techniques estimate the same values of mean arrival rates because they both use overload-underload approach. For

| Matching Technique | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ ($s^{-1}$) | $r_2$ ($s^{-1}$) |
|---|---|---|---|---|
| Moment matching | 6670.2 | 5155.4 | 2.1858 | 1.8767 |
| $\sum$-matching | 6651.7 | 5359.6 | 20.9298 | 13.0209 |
| IDC matching | 6651.7 | 5359.6 | 1.7333 | 1.0783 |

Table 3.5: The equivalent MMPP parameters by using different matching techniques

moment matching and IDC matching, who both match IDC, the values of mean transition rates are close.

Now let us examine each of the proposed MMPP models and compare them to the reference aggregated on-off sources. In Figure 3.7 the IDC curves for all three MMPP models and also the reference aggregated on-off sources have been shown. The moment-based matching gives the most accurate estimation of IDC curve, expectedly, as it matches the IDC curve at two points. However, it is impossible to fully match the IDC curve for on-off source and MMPP, because for MMPP the value of IDC at very small lags ($t \to 0$) approaches 1 while for on-off source it approaches zero. IDC matching technique gives a matching at IDC($\infty$). As we showed in the previous sections, having the same IDC($\infty$) gurrantees the same slope for IDC curve in the rising region too. $\sum$-matching does not macth the IDC curve. As a matter of fact, the value of IDC for the model deriving by this technique is closer to that of Poisson (IDC = 1) than on-off source.

We then simulated a G/D/1 queue with each of the above traffic models as the source, in order to compare the queueing performance and to see how well each of the techniques can

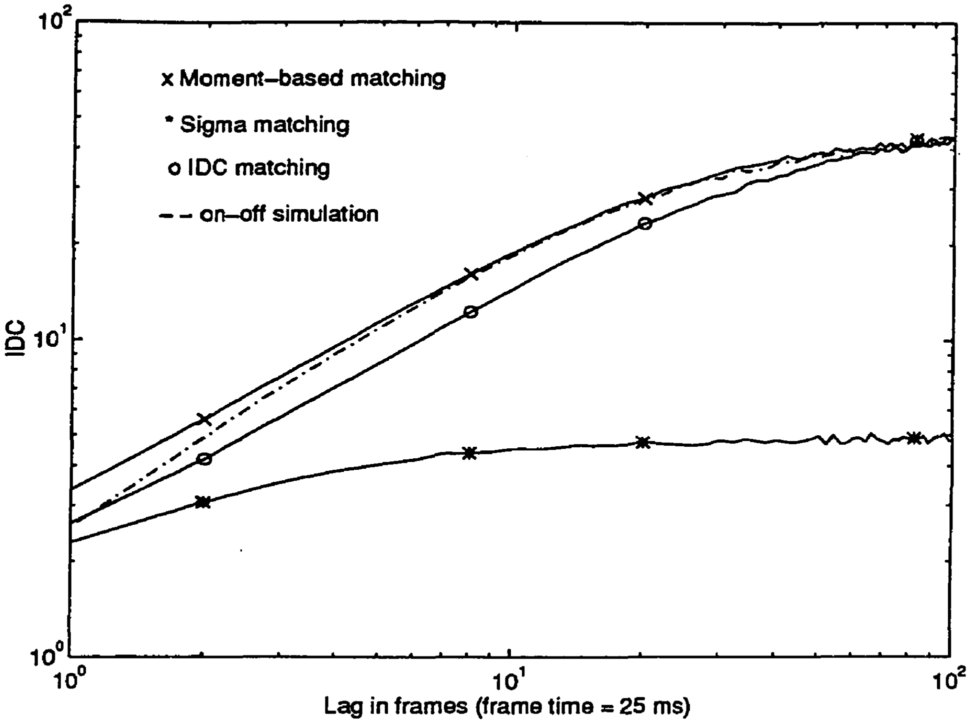

Figure 3.7: IDC curves for different on-off/MMPP matching techniques

predict the average delay and the probability of loss. We used *OPNET* network simulator for this study. In Figures 3.8 and 3.9 the average delay vs. traffic load and probability of loss vs. buffer size have been shown. One point here is noteworthy. As we explained in Chapter 2, the curve of the probability of loss vs. buffer size has been approximated by the curve of the survivor function of the queue length vs. buffer size for an infinite buffer case. This is the approximation which we use throughout this thesis wherever we talk about the probability of loss. The corresponding curves for an M/D/1 queue with the same load have been presented for comparison.

In the Figure 3.8, the value of delay has been normalized by the service time. The results indicate that the three matching techniques provide the same performance in low traffic loads. At loads higher than 0.75, the $\sum$-matching technique gives an optimistic result

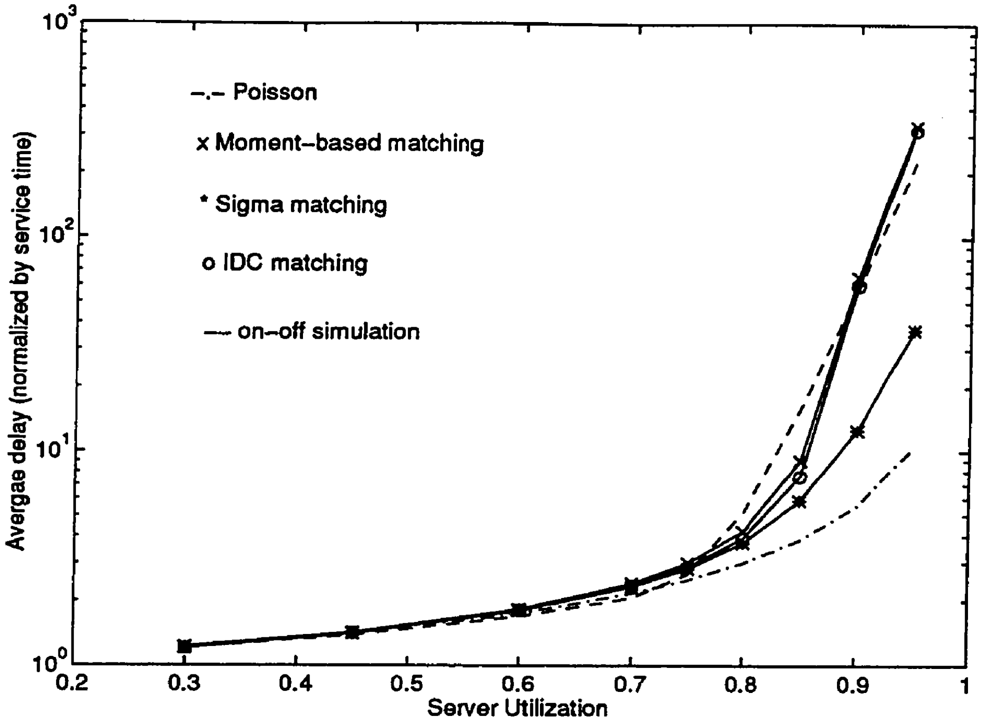

Figure 3.8: Delay vs. load curves for different on-off/MMPP matching techniques

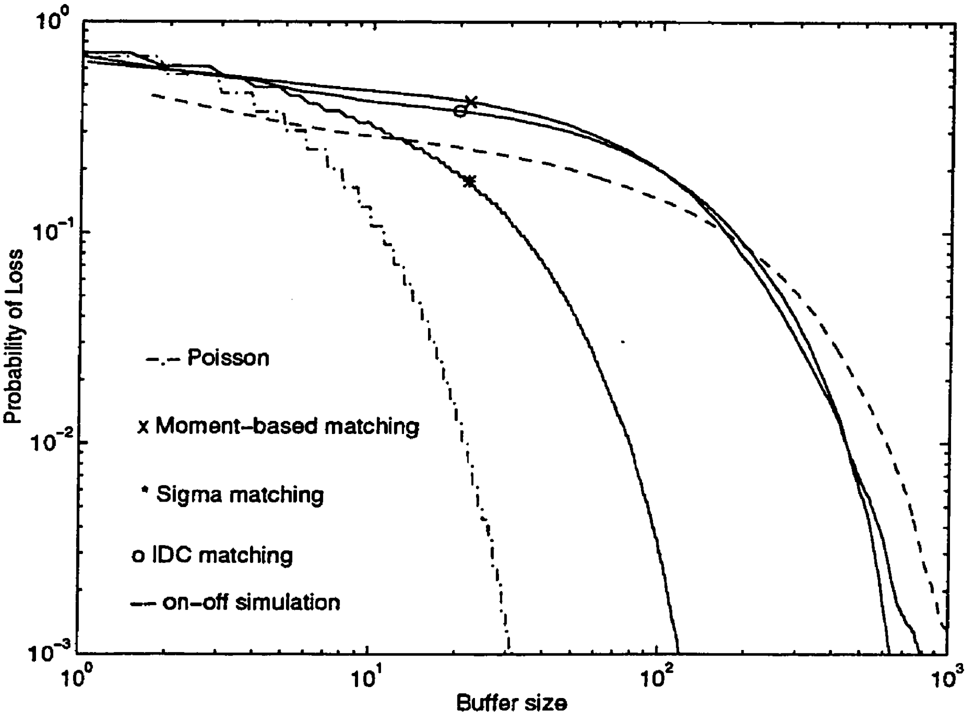

Figure 3.9: Survivor functions for different on-off/MMPP matching techniques

closer to the Poisson model than to the aggregated on-off voice sources. The results of the moment-based and IDC matching techniques are very close and in a good agreement with those of aggregtaed on-off voice sources.

Figure 3.9 shows the loss rate versus buffer size for a traffic load of 0.9. The results indicate a good agreement between the the performance of the aggregated on-off voice sources and its equivalent derived by the moment-based and IDC matching techniques, while the $\sum$-matching technique gives a rather optimistic loss rate.

The above results indicate the important role of the IDC in parameter matching. Furthermore, while a moment-based matching gives the closest results, a much simpler overload-underload IDC matching could be used to match a 2-state MMPP to a superposition of on-off sources effectively, in order to avoid the complexity and lengthy calculations of inverse transforms in the moment-based matching technique.

## 3.4 Modeling of an arbitrary traffic and parameter estimation

In this section we are going to generalize the modeling technique for an arbitrary traffic, not a special case of the aggregated voice sources. First we start with a generalized moment-based matching technique and then we will review the other proposed techniques.

### 3.4.1 Generalized moment-based matching

Both the $\sum$-matching and IDC matching techniques are based on the overload-underload approach and hence on the assumption that the traffic is a superposition of on-off sources. For this reason, they are not suitable to model an arbitrary traffic. In this section we will focus on the moment-based technique.

With only the first and second moments (or IDC), the four parameters of the equivalent 2-state MMPP model cannot be uniquely determined. Therefore, for an arbitrary traffic we need to use the third moment. In [17] the parameter estimation is based on the measurement of one or two points of input parameters (IDC and the third moment). However, the selection of these points has a large impact on the performance results because of the error resulting from the limited number of samples. Consider a simple estimation using the ensemble mean from n samples $s_1, s_2, \ldots, s_n$, i.e., $\frac{1}{n}\sum_{i=1}^{n} s_i$. An accurate estimation requires a sufficiently large values of n and hence a large observation time interval. Otherwise, some measured samples with a large variation can greatly influence the captured mean. In the following procedure, we suggest a filtering approach suitable to an accurate estimation without requiring a large observation interval.

- Establish the histogram of samples using an arbitrary number of bins based on the maximum and minimum value of the samples
- If the peak of the histogram is less than 0.5, decrease number of bins by a factor of 2 and modify the histogram
- The process is repeated over until:
  - The peak of the histogram is more than 0.5, or
  - The number of bins is 2 (i.e. the peak will be at least 0.5).
- Select only the samples in the peak bin to compute the sample mean

Figure 3.10 shows how the technique works. It starts with 40 equally-spaced bins and computes the histogram, then reduces the number of bins and continues. At bin number=5,

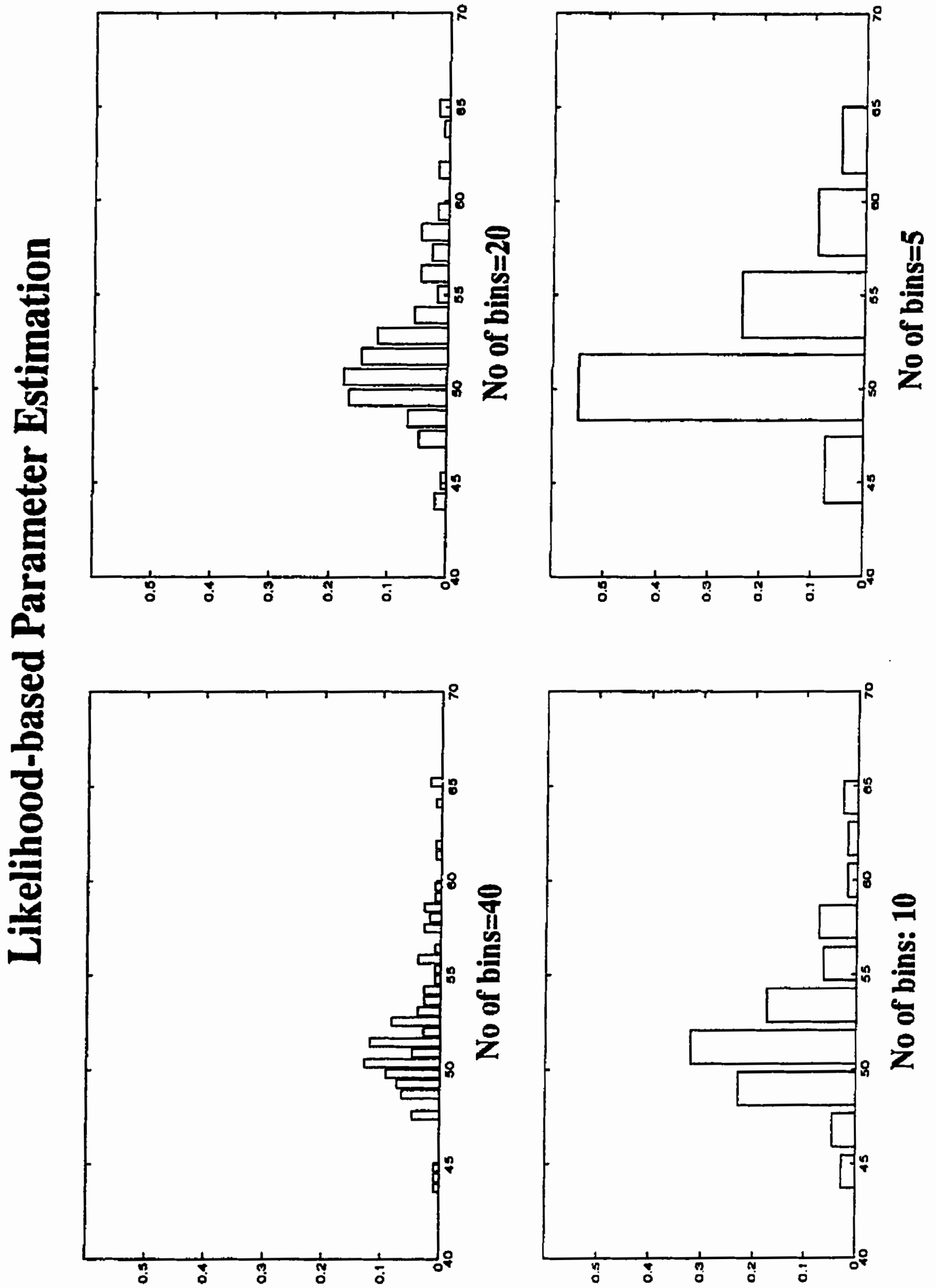

Figure 3.10: The likelihood-based parameter estimation

the peak of the histogram has passed 0.5, so then the samples will be filtered and only those who reside in the area specified by the peak bin in Figure 3.10 will be picked.

This estimation technique is used in our matching to compute the values of $IDC(\infty)$, $d = r_1 + r_2$, and $\hat{\lambda} = \frac{\lambda_1 r_1 + \lambda_2 r_2}{r_1 + r_2}$ which is a temporary parameter introduced in the technique to simplify the matching process. $\hat{\lambda}$ can be calculated from the third moment of the samples using the following equation [27]:

$$\hat{\lambda} = \{\mu_3(0:t) - (\frac{6Ia^2}{d^2} - \frac{3aI}{d})(1 - e^{-dt}) + \frac{3a^2It}{d}(1 + e^{-dt}) - 3atI - at\} / \qquad (3.28)$$

$$\{\frac{3\,a\,It}{d}(1 + e^{-dt}) - \frac{6\,a\,I}{d^2}(1 - e^{-dt})\}$$

where:

$$a = \text{mean arrival rate} = \overline{\lambda} \;, \quad d = r_1 + r_2 \qquad (3.29)$$

$$I = IDC(\infty) - 1 \;, \quad \hat{\lambda} = \frac{\lambda_1 r_1 + \lambda_2 r_2}{r_1 + r_2}$$

Using $\hat{\lambda}$ and the parameters defined in(3.29), we compute [27]:

$K = d\,(\hat{\lambda} - a)$,

- If $K = 0$ , $r_1 = r_2 = d/2$

  $\lambda_1 = a\,(1 + \sqrt{dI})$ , $\lambda_2 = a\,(1 - \sqrt{dI})$

- If $K \neq 0$, $e = \frac{ad^3 I}{2K^2}$

  $r_1 = \frac{d}{2}(1 + \frac{1}{\sqrt{4e+1}})$ , $r_2 = d - r_1$

  $\lambda_2 = (\frac{a\,d}{r_2} - \frac{K}{r_1 - r_2})\frac{r_2}{r_1 + r_2}$ , $\lambda_1 = \lambda_2 + \frac{K}{r_1 - r_2}$

Table 3.6 shows the estimated values of the parameters of the model for 2-state MMPP traffic with known parameters and generated by simulation. The test cases were chosen carefully to cover a range of different $r_1/r_2$, $r_1 + r_2$ and $IDC(\infty)$. The accuracy of the technique is noticable. For mean arrival rates at each state, the error is less than 1% in all of the cases. For transition rates it is less than 7%.

| Cases | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ (1/s) | $r_2$ (1/s) |
|---|---|---|---|---|
| Original | 6651.7 | 5359.6 | 1.7333 | 1.0783 |
| Estimated | 6644.5 | 5360 | 1.6298 | 1.0321 |
| Original | 4233.4 | 2239.9 | 3.1522 | 1.9425 |
| Estimated | 4224.4 | 2235 | 3.0747 | 1.9616 |
| Original | 4574.6 | 2785.8 | 2.4041 | 0.3271 |
| Estimated | 4541.7 | 2782 | 2.2024 | 0.3467 |
| Original | 3271.6 | 2728.4 | 0.6514 | 0.6514 |
| Estimated | 3279.1 | 2727.1 | 0.6724 | 0.6908 |

Table 3.6: Test cases for refined matching technique

### 3.4.2 Other techniques

There are a number of other techniques which could be used for fitting a 2-state MMPP model to an arrival process. Here we briefly review a couple of these techniques.

#### 3.4.2.1 Gusella's method

Gusella in [25] proposed a moment-based technique for fitting a 2-state MMPP model to an arrival process. The difference between his technique and ours lays in the point that he uses the ***squared coeeficient of variation*** of the interarrival times instead of the third central moment of the counting process. The following steps have to be taken in his technique for deriving the parameters:

1. From the data samples, estimate $a$ the mean arrival time, $b$ the limiting value of IDC minus 1 (IDC($\infty$) $-$ 1), and $d$ the squared coefficient of variation of the interarrival times.

2. Compute the value of IDC at an arbitrary lag, $t_0$. Using $b$ and $I_{t_0}$ (the value of IDC at $t_0$), find the mean rate $c = r_1 + r_2$ in the following equation:

$$\frac{b - I_{t_0}}{b - 1} = \frac{1 - e^{-ct_0}}{ct_0} \tag{3.30}$$

3. Obtain a value for $\lambda_2$ from the following equation:

$$d = \frac{2a\lambda_2^2 + (2ac + abc - 2)\lambda_2 - 2c(b + 1)}{2a\lambda_2^2 + (2ac + abc - 2)\lambda_2 - 2c} \tag{3.31}$$

Then find the other parameters as follow:

$$\begin{aligned}\lambda_1 &= \tfrac{2+abc-2a\lambda_2}{2a-2a^2\lambda_2}\\ r_1 &= \tfrac{abc^2}{2+abc-4a\lambda_2+2a^2\lambda_2^2}\\ r_2 &= \tfrac{2c(a\lambda_2-1)^2}{2+abc-4a\lambda_2+2a^2\lambda_2^2}\end{aligned} \tag{3.32}$$

4. Based on the current values of the parameters of the MMPP model, compute the goodness of the approximation by comparing the estimated IDC with the theoretical one which is calculated from Equation (3.1). A minimum squraed error test could be used. Adjust the value of c and repeat steps 3 and 4 until a satisfactory approximation is reached.

Gusella's method is in nature moment-based, so not much different from our generalized moment-based matching technique. However, the technique needs measurements from both time and counting processes which may be difficult particularly in the simulations.

#### 3.4.2.2 Likelihood-based technique

For a detailed account of this class of techniques refer to [20]. Here we just briefly discuss a technique introduced by *Meier-Hellstern* for fitting a 2-state MMPP model to the arrival process.

Meier-Hellstern's likelihood-based technique [19] computes the matrices of the model, defined in Equations (2.7) and (2.8) from the samples of interarrival time. One interesting point in Meier-Hellstern work is her note that if the data is close to having a Poisson model, then any MMPP fitting technique may fail because the Poisson model is normally a superposition of infinite number of independent processes. Therefore she recommends a Poissonness test at the beginning. The following steps are taken in the algorithm [19].

Input parameters:

- $\{x_i\}_{i=1}^n$: Observed interarrival time sequence from the arrival process

- $Q$: Mean transition rate matrix for the MMPP model

- $\Lambda$ : Mean arrival rate matrix for the MMPP model

- $L$: Likelihood function for $Q$ and $\Lambda$ given the observed interval sequesnce $x_{i}{}_{i=1}^n$. The function $L$ for 2-state MMPP is derived as:

$$L(Q, \Lambda \,|\, x_1, \ldots, x_n) = \pi \Pi_{k=1}^n \{\exp[(Q - \Lambda) x_k] \Lambda\} e \tag{3.33}$$

- $J_k$: The *phase* of the Markov process at $k_{th}$ interval

The set of the parameters of the 2-state MMPP model, $[\lambda_1\ ,\ \lambda_2\ ,\ r_1\ ,\ r_2]$ is replace by the following alternative set:

- $\pi_1 = r_2(r_1+r_2)^{-1}$, the stationary probability of being in state 1 of the time-stationary MMPP. Therefore the steady state probability matrix will be $\pi = [\pi_1\ \ ,\ \ 1 - \pi_1]$

- $\lambda^* = \lambda_1\pi_1 + \lambda_2(1 - \pi_1)$, the mean arrival rate for 2-state MMPP.

- $P_{11} = \lambda_1(\lambda_2 + r_2)(\lambda_1\lambda_2 + \lambda_1 r_2 + \lambda_2 r_1)^{-1}$, the probability of a transition from state 1 to state 2.

- $p_1 = \lambda^{*-1}\lambda_1\pi_1$, the steady-state proportion of arrivals from state 1.

If $\lambda_2 \neq 0$, these formulas establish a one-to-one correspondence between the new parameters and $\lambda_1$, $\lambda_2$, $r_1$, and $r_2$. The model parameters can be calculated from the alternative parameters by using the following formulas:

$$
\begin{aligned}
\lambda_1 &= \lambda^* p_1 \pi_1^{-1} \\
\lambda_2 &= \lambda^*(1-p_1)(1-\pi_1)^{-1} \\
r_1 &= \lambda_1(1-p_1)(1-P_{11})(P_{11}-p_1)^{-1} \\
r_2 &= \lambda_2(1-P_{11})(P_{11}-p_1)^{-1}
\end{aligned}
\tag{3.34}
$$

The Algorithm:

1. Estimate the mean of the interarrival time from the samples of the arrival process. The mean arrival rate will be calculated as $\hat{\lambda}^* = n(\sum_{i=1}^{n} x_i)^{-1}$.

2. Test for Poissonness of the data. If the observed stream is statistically indistinguishable from a Poisson process of rate $\hat{\lambda}^*$, then stop the algorithm and use the Poisson model.

3. Construct an initial estimate $(J_k^{(0)})_{k=0}^{n}$ of the sequence $(J_k)_{k=0}^{n}$. First smooth the data using a moving average scheme based on the arithmetic mean. Classify the elements of $(J_k)_{k=0}^{n}$ as 1 or 2 depending on whether the smoothed interarrival times are greater or less than $\frac{1}{\hat{\lambda}^*}$.

4. Set $r = 0$, $V^{(r)} = 0$.

5. Let:

$$
\begin{aligned}
\hat{p}_1^{(r)} &= n^{-1} \sum_{k=1}^{n} I(J_k^{(r)} = 1) \\
\hat{P}_{11}^{(r)} &= n_{11}(n_{11} + n_{12})^{-1}
\end{aligned}
\tag{3.35}
$$

   where $I$ denotes the indicator function and $n_{ij} = \sum_{k=2}^{n} I\left[J_k^{(r)} = j \;\middle|\; J_{k-1}^{(r)} = i\right]$.

6. Let $\hat{\pi}_1^{(r)}$ be the maximum likelihood estimation of $\pi_1$ given $\hat{\lambda}^*$, $\hat{p}_1^{(r)}$, $\hat{P}_{11}^{(r)}$, $\{J_k^{(r)}\}_{k=0}^{n}$,

and $\{x_k\}_{k=1}^n$. Then for $0 < \pi_1 < 1$, the matrices $\Lambda$ and $Q$ may be considered as functions of $\pi_1$, based on Equations (3.34). We use the notations $Q^{(r)}(\pi_1)$ and $\Lambda^{(r)}(\pi_1)$ for these functions.

The likelihood function is defined as:

$$\log L^{(r)}(\pi_1) = \sum_{k=1}^{n} \log F'_{J_k^{(r)}-1, J_k^{(r)}}(r, \pi_1; x) \tag{3.36}$$

where

$$\begin{aligned} F'(x) &= \left[F'_{ij}(x)\right] = \exp\left[(Q - \Lambda)x\right]\Lambda \\ F'(r, \pi_1; x) &= F'(x)\Big|_{Q=Q^{(r)}(\pi_1), \Lambda=\Lambda^{(r)}(\pi_1)} \end{aligned} \tag{3.37}$$

7. Let $\pi_{(r)}$ be the value of $\pi_1$ which maximizes $\log L^{(r)}(\pi_1)$ using a numerical maximization technique. Let $Q = Q^{(r)}(\pi_1^{(r)})$, $\Lambda = \Lambda^{(r)}(\pi_1^{(r)})$.

8. Let:

$$V^{(r+1)} = \max_A \sum_{k=1}^{n} \log F'_{J_{k-1}, J_k}(r; x) \tag{3.38}$$

where

$$A = \{(J_0, J_1, \cdots, J_n) \,|\, 1 \leq J_k \leq 2, \; 0 \leq k \leq n\}$$

and

$$F'(r; x) = F'(x)\big|_{Q=Q^{(r)}, \Lambda=\Lambda^{(r)}}$$

and $\left[J_k^{(r+1)}\right]_{k=0}^{n}$ is defined to be the sequence which maximizes (3.38).

9. If $|\, V^{(r+1)} - V^{(r)} \,| \leq \delta$ for sufficiently small values of $\delta$, set $r = r + 1$ and go to step

#6.

For more details and remarks on the technique as well as study of behaviour and performance of the likelihood-based technique, refer to [19].

Now back to our generalized moment-based technique, it enables us to derive a model from traffic samples. Let us introduce a technique to predict the queueing performance of the model in order to make it easy to use in traffic control algorithms.

### 3.5 Performance approximation of MMPP/D/1 queue

The MMPP/D/1 queue could be solved by using the matrix geometric technique introduced in Section 2.8.1. However, this complex, iterative technique needs a lot of computation power and time. For real-time traffic control we are going to need an approximation which can be calculated quickly and provides us with enough accuracy. Here we present such an approximation [27].

The closed form expression for the Laplace-Stieltjes transform of delay for MMPP/G/1 queue is [18]:

$$D(s) = s\,(1-\rho)\,g\,[sI + Q - \Lambda\,(1 - H(s))]^{-1}\,e \tag{3.39}$$

Where Q and $\Lambda$ are the infinitesimal generator and arrival rate matrices for MMPP model, respectively, H(s) is the Laplace-Stieltjes transform of the service time and $\rho$ denotes the utilization. $g=[g_1 \;\; 1-g_1]$ is determined by solving numerically the following equation for $g_1$:

$$\begin{aligned}&\frac{r_1 - g_1\,(r_1 + r_2)}{\lambda_1 - \lambda_2}\,(\frac{1}{g_1} + \frac{1}{g_1 - 1}) \\ &+\exp[\frac{(1-g_1)\,\lambda_1\,r_1 - g_1\,\lambda_2\,r_2}{(1-g_1)\,g_1\,(\lambda_1 - \lambda_2)}] = 1\end{aligned} \tag{3.40}$$

For MMPP/D/1 queues, $H(s) = e^{-s\,h}$ where h denotes the cell service time. If we assume that h is small enough as compared to average delay, we can safely ignore the incomplete service time upon cell arrival. In this case, the delay is the product of the number of cells in queue and the cell service time. Consequently, the probability of buffer overflow can be calculated from delay survivor function as $\Pr(X > t/h)$=$\Pr(T > t)$. In other words, the same equation as (3.39) can also be used for Laplace-Stieltjes function of the queue length, at least for burst region.

Following the same approach taken by [31] to estimate delay's survivor function, we approximate the probability of buffer overflow by a single exponential function $\alpha\, e^{s_1\,h\,x}$, in which $s_1$ is the largest negative root of the denominator of (3.39) (the closest negative root to zero. Here we simplify the approach in [31] further by approximating the exponential form of H(s) with the first three components of its Maclaurin series, $1 - h\,s + \frac{1}{2}\,(h\,s)^2$. Substituting it into (3.39), $s_1$ is the largest negative root (the closest negative root to zero) of the following cubic equation:

$$\frac{1}{4}\lambda_1\lambda_2 h^4 s^3 + h^2[\frac{1}{2}(\lambda_1+\lambda_2) - \lambda_1\lambda_2 h]\,s^2 + \qquad (3.41)$$

$$[(1-\lambda_1 h)(1-\lambda_2 h) - \frac{1}{2}\rho r_1 r_2 h]\,s + (r_1+r_2)(\rho-1) = 0$$

in which $\rho = \overline{\lambda}\,h$. We approximate the queue length by system size (for the burst region), so $\alpha = \rho$ and the probability of buffer overflow is computed as:

$$\Pr(X > x) = \rho\, e^{s_1\,h\,x} \qquad (3.42)$$

Figure 3.11 shows a good agreement in the results using simulations and the approximation given by Equation (3.42) for an MMPP/D/1 queue. Equation (3.42) can be a simple tool for buffer design.

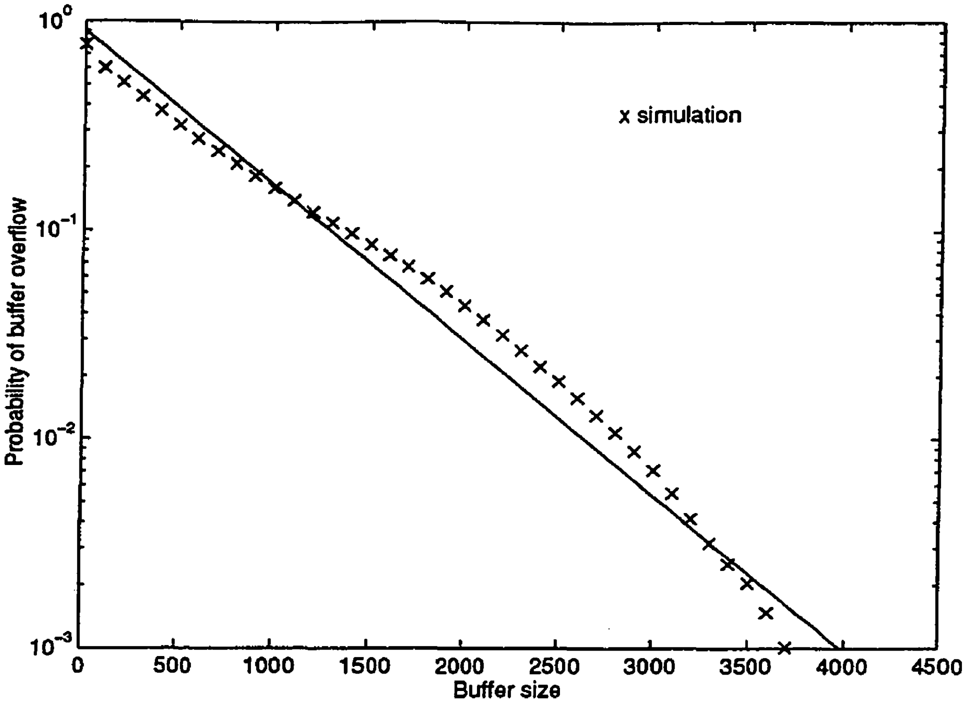

Figure 3.11: The accuracy of the approximation of the loss

Our simulations shows that the approximation for the exponential slope of survivor function is accurate enough in almost every case. However, there are some cases where the approximation for the coeficient ($\rho$) is not precise enough. We assumed that the system has a very short cell level. In the cases where the cell-level region is too long or the probability of loss in this region drops rapidly, the value of probability of loss given by Equation ( 3.42) may be a bit higher than what we get by using simulation.

### 3.6 Shortcomings of the 2-state MMPP model

The 2-state MMPP model has been widely used for its simplicity and analytical tractability. However, there exist some cases where the model is unable to represent the effects of the high burstiness of the aggregate multimedia traffic. In these applications, the model usually underestimates the probability of loss for a given buffer size, so cannot keep up with the long tail of the loss curve. This problem mainly is due to the fact that a too-simple 2-state MMPP does not have the flexibility to model all various *phases* in the multimedia traffic.

One solution is to increase the number of states and to upgrade our model from a 2-state to a multiple-state MMPP. Many problems arises in this modification, including a sharp increase in the number of parameters which must be determined in the matching process, the complexity of the model as well as the time and computation power needed for solving the system analytically. In the next chapter we will discuss the multiple-state MMPP models, and introduce a simple multiple-state MMPP that enjoys the advantages of the 2-state MMPP (simplicity and analytical tractibility) while is capable of representing a higher number of phases in the traffic in order to get a more precise prediction of the queueing performance.

# CHAPTER 4
# Modeling of Aggregate ATM Traffic using multiple-state Makov Modulated Poisson Processes

## 4.1 Introduction

In the previous chapters we showed that the 2-state MMPP is a good model for aggregate voice sources [17] and analytically tractable for its simplicity. However, it may not be able to represent the effects of the high burstiness of aggregate multimedia traffic. For example, an aggregate multimedia traffic stream at the input port of an ATM switch switch can be considered as a superposition of three components: voice, video and data. If each traffic component is modeled by a 2-state MMPP, then the superposition will be a multi-state MMPP [18].

The increase in the number of states causes an increase in burstiness of the traffic and therefore the queue length or probability of loss in the multiplexers will also increase. In the next section we will offer an example to show that. Due to the above fact, there are numerous studies on the possibility of using multiple-state MMPP models to represent the long range dependent traffic [33] [35]. 2-state MMPP models usually underestimate the probability of loss in ATM multiplexers with the long-range dependent traffic at the inputs. Therefore the multiple-state MMPP are used to represent those traffics.

In the coming sections, we first study the general problem of representing an ATM traffic by a multiple-state MMPP model, and we present our model which is a special case of mutiple-state MMPP. We discuss various techniques to derive the parameters of such model, and then we study its performance by applying it in a number of case studies.

### 4.2 Multiple-state MMPP for ATM multimedia Traffic

In general, a multiple-state MMPP model is identified by two matrices, a cell generation rate matrix $\Lambda$ and a transition rate (or infinitesimal) matrix $Q$. For an N-state MMPP these matrices are $N \times N$. For the transition rate matrix, one of the elements on each row can be calculated from the values of others. Therefore a general N-state MMPP has $2N^2 - N$ different parameters. There have been some efforts to derive the complete matrices from the samples of the traffic. However the techniques are often complicated and give inconsistent results as we explain it in the next section.

To overcome this problem, we decided to use a simpler case. We use a special case of multiple-state MMPP, a superposition of 2-state homogenous MMPP minisources. The advantages of this model are as follows:

- The model has only five parameters: Number of minisources, N, and four parameters of a minisource, $\lambda_1$ and $r_1$, $\lambda_2$ and $r_2$, which are the respective cell generation rate and transition rate in each of the states. In fact this model has only one parameter more than a simple 2-state MMPP, which is the number of minisources, N.
- In general, the superposition of N 2-state MMPP results in a $2^N$-state MMPP model. But for this special case, the number of states reduces to N+1 [18].
- Both the cell generation rate and transition rate matrices can be computed directly without using *Kronecker* summation [18]. The computations are simpler for this

special case.

The infinitesimal matrix $Q$ and cell generation rate matrix $\Lambda$ can be calculated for our model as follows [18]:

$$\Lambda = \mathrm{diag}(j\,\lambda_1 + (N-j)\,\lambda_2) \quad j = 0 : N \tag{4.1}$$

$$\begin{array}{ll} Q(j,j) = -j\,r_1 - (N-j)\,r_2 & \\ Q(j,j+1) = (N-j)\,r_2 & j = 0 : N \\ Q(j,j-1) = j\,r_1 & \\ & \\ Q(j,j+i) = 0 & |i| > 1 \end{array} \tag{4.2}$$

Now let us see how we can determine the parameters of this model for a given sequence of traffic samples.

## 4.3 Estimation of the parameters of Multiple-state MMPP model from traffic samples

We developed a Pdf-based technique for deriving the parameters of multiple-state MMPP model from empirical data. But before describing the technique in detail in next section, here let us review the shortcomings of moment-based technique in mutiple-state cases along with a few other techniques.

### 4.3.1 Shortcomings of Moment-based technique in multiple-state case

In previous chapter we presented a moment-based technique for estimating the parameters of a 2-state MMPP model. Here let us explain why this technique cannot be applied in multiple-state MMPP case.

In our moment matching technique, we used moments of the counting process of the

| Parameter | Source 1 | Source 2 |
|---|---|---|
| N | 1 | 6 |
| $\lambda_1$ | 3652.1 Cells/s | 800 Cells/s |
| $\lambda_2$ | 2447.9 Cells/s | 300 Cells/s |
| $r_1$ | 2.1661 ($s^{-1}$) | 2.4 ($s^{-1}$) |
| $r_2$ | 1.8339 ($s^{-1}$) | 1.6 ($s^{-1}$) |

Table 4.1: MMPP models with the same moments

traffic samples to estimate the unknown parameters. However, it is not difficult to show that a 2-state MMPP can have the same moments as a multiple-state MMPP but with different queueing performance. In particular, for our model of the superposition of N 2-state homogenous MMPP minisources, it is easy to show that the following relations exist:

$$\mathrm{IDC}_N(t) = \mathrm{IDC}(t) \tag{4.3}$$

$$\mu_N^3(t) = \mu^3(t) \tag{4.4}$$

where $\mathrm{IDC}_N(t)$ and $\mu_N^3(t)$ denote the IDC curve and third central moment of the aggregated model, and $\mathrm{IDC}(t)$ and $\mu^3(t)$ the respective parameters of a minisource. It is even possible to build two models with exactly the same mean arrival rate, IDC curve and the third central moment, but with very different queueing performance. As an example, consider two sources: one represented by a single 2-state MMPP and the other by a superposition of N 2-state MMPP's (equivalent to an N+1-state MMPP) with the following parameters:

It can be verified by using the above parameters and Equations (3.2) and (3.9) along with Equations (4.3) and (4.4), that both of the sources have exactly the same first three moments for a counting process: average, IDC curve and third central moment. The Equations Figure 4.1 shows the probability of loss (or buffer overflow) versus buffer length of an

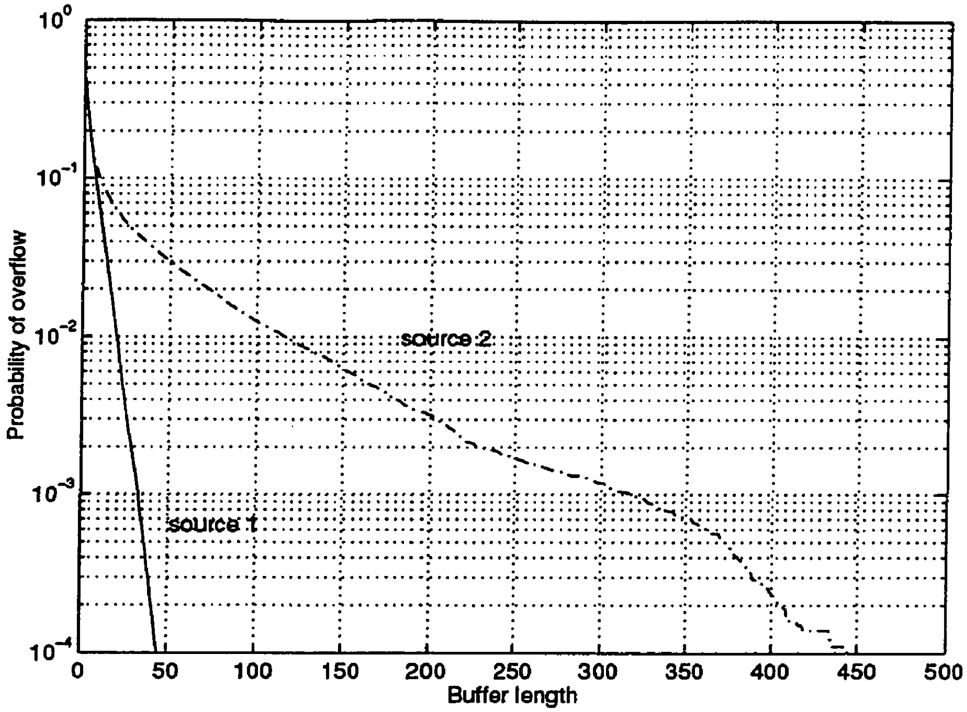

Figure 4.1: Effect of number of states on queueing performance

MMPP/D/1 queue for both cases at a traffic load of 0.75 . The curves indicate a huge difference in queueing performance between two sources, though they have the same first three moments.

The above results show that the moments of the counting process do not contain enough information about the burstiness or number of states. Therefore the moment-based technique we applied in previous chapter for modeling of an arbitrary traffic by a 2-state MMPP is not applicable for a general case with multiple states. In fact, it is possible to find an equivalent-in-moments 2-state MMPP model for a superposition of N 2-state MMPP minisources. The procedure is simple. Suppose that we have a superposition of N homogenous 2-state MMPP minisources, each with parameters $\lambda_1$ and $r_1$, $\lambda_2$ and $r_2$, the cell generation rate and state transition rate in each state, respectively. Then we want to build another 2-state MMPP model, with parameters $\hat{\lambda_1}$, $\hat{r_1}$, $\hat{\lambda_2}$ and $\hat{r_2}$ so that both processes have exactly the same mean, IDC curve and third central moment. We follow theses steps:

- The mean arrival rate of the aggregated traffic is calculated by using the following general formula:

$$\overline{\lambda} = \frac{N\,(\lambda_1 r_2 + \lambda_2 r_1)}{r_1 + r_2} \tag{4.5}$$

- The IDC curve and the third central moment for aggregated traffic are calculated using Equations (3.2) and (3.9) and considering Equations (4.3) and (4.4).

- Now to build the 2-state model with parameters $\hat{\lambda_1}$, $\hat{r_1}$, $\hat{\lambda_2}$ and $\hat{r_2}$ with the above mean, IDC and third central moment, we follow the same moment-based technique we used in section 3.4.1 for estimating the parameters of a 2-state MMPP model from its moments.

The above results show that a purely moment-based technique cannot be used to derive the parameters of a multiple-state MMPP model.

### 4.3.2 Histogram-based technique

A histogram-based technique for estimation of the parameters of a multiple-state MMPP from empirical data was proposed by Skelly et al in [21]. The authors used this technique for modeling of video traffic behaviour in ATM multiplexers.

The technique is purely histogram-based. First of all, the user chooses an arbitrary number of states. A number of bins of eight has been recommended in the paper. Then the traffic sequence is quantized to correspond to the allowable arrival rates. Each of these allowable rates is corresponding to a state of MMPP model.

In the next step, the transition probabilities are measured from the empirical data. When the frame period is deterministic, the transition rate matrix may be computed from the transition probabilities matrix by applying the following equation:

$$Q = f(P - I) \tag{4.6}$$

in which Q denotes the transition rate matrix, P is the transition probability matrix, I is identity matrix and f denotes frame rate.

The authors have reported agreement between the results of this technique with measured video traffic. However we found that the technique fails to estimate the transition rates for an MMPP source, even in 2-state case which is the simplest. We used the sample results of the simulation of MMPP sources with known parameters. We believe that the reason behind the failure of the histogram-based technique for estimation of the parameters of an MMPP model lays in the fact that the quantization of the arrival rates in the case of MMPP traffic results in a poor accuracy. In each state of a multiple-state MMPP model, the number of arrivals over an interval follows Poisson process and so may accept any positive value from zero upto infinity. Therefore, having a specific number of arrivals during a frame period does not specify in which state the process sojourns. In other words, the idea of quantization in the case of MMPP is meaningless. Consequently, there is no accurate way to measure the transition probabilities directly from the empirical data.

### 4.3.3 Likelihood-based techniques

A good review of these techniques can be found in [20]. The idea is to employ a maximum likelihood estimation to find transition rate and cell generation rate matrices from empirical data. An example of these techniques is the one by *Meier-Hellstern* [19] which we briefly described in Section 3.4.2. This technique, which uses samples of interarrival times for estimation of model parameters, is applicable to multiple-state case too, but as we mentioned before, it may give inconsistent results and needs a good initial point [20]. We do not go into the details for this technique as our concentration is on the techniques which use the counting process, not time process.

In [36] a recursive, likelihood-based technique has been proposed. The technique is applicable for base cases of samples of the counting process or time process. Although, for counting process to give good results, the sampling rate must be much higher than state transition rates. In other words, the observation interval must be short enough. We explain more about this limitation when we introduce our pdf-based technique in the next section. This recursive technique uses the conditional state transition probabilities given the number of arrivals in a frame or given the interarrival time. However, in this technique the results depend heavily on the initial point as in Meier-Hellstern technique.

After studying various measures which characterize the traffic, we came to this conclusion that two traffic measures contain sufficient information to uniquely identify the process: the probability density function (pdf) of arrival rate, and the index of dispersion for counts (IDC) curve that captures the correlation effect. Now here we will introduce a new pdf-based technique which uses these two parameters of the traffic to estimate the parameters of a multiple-state MMPP model for the traffic. The technique is quite easy to implement, does not need any initial guess about the parameters, very consistent and estimates the model parameters pretty good.

## 4.4 Pdf-based matching technique

### 4.4.1 Model Parameters

We explained that we want to model ATM multimedia traffic by a model represented by a superposition of N independent and homogenous 2-state MMPP minisources. This proposed model is fully described by five parameters [$\lambda_1$, $\lambda_2$, $r_1$, $r_2$, $N$]. They are estimated from the probability density function (or histogram) of number of arrivals over an observation interval, and the curve of index of dispersion for counts (IDC) of the samples.

The IDC curve for a single 2-state MMPP minisource can be calculated using Equation (3.2). The IDC curve of the superposition of identical and statistically independent processes is the same as the IDC curve of each individual process as Equation (4.3) shows. So that we can estimate two parameters, $d$ and $IDC(\infty)$ from the IDC curve based on the traffic samples. To make our matching technique easier, we derive the following alternative set of parameters for our model:

$$[\ \overline{\lambda}\quad d\quad IDC(\infty)\quad \alpha\quad N]$$

where $\alpha = r_2/(r_1 + r_2)$ is the steady-state probability of staying at state 1, and $\overline{\lambda}$ = mean arrival rate. The original parameters could be calculated from the above alternative set as follows:

$$r_2 = \alpha\, d \qquad r_1 = d - r_2$$

$$\lambda_1 = \frac{\overline{\lambda}}{N} + \sqrt{\frac{\frac{1}{2}\, d\, \overline{\lambda}\, (IDC(\infty) - 1)\ (1-\alpha)}{N\, \alpha}} \tag{4.7}$$

$$\lambda_2 = \frac{\overline{\lambda}}{N} - \sqrt{\frac{\frac{1}{2}\, d\, \overline{\lambda}\, (IDC(\infty) - 1)\ \alpha}{N\, (1-\alpha)}} \tag{4.8}$$

Here $\lambda_1$, $\lambda_2$, $r_1$ and $r_2$ are the parameters of the *minisource model.*

The next section outlines the procedure to derive the five parameters of alternative set from samples of the traffic.

### 4.4.2 Derivation of the probability density function

Consider a single 2-state minisource. The probability density function for counting process can be analytically derived using a complex iterative procedure [18]. In the following, we propose a much simpler approximation.

In general, we can use any observation interval for counting the number of arrivals. However, if the selected interval is small enough compared to the sojourn time at each state, then there will be no state change during an observation inetrval T. In other words, we assume that we can effectively measure the *rate* of arrivals at each epoch (a short observation interval). This assumption of *arrival rate* instead of *number of arrivals over an observation interval* is reasonable in practice and confirmed by our simulation results, to be shown later.

If no state change happens during an observation interval, the pdf of arrival rate can be calculated using the conditional pdf's as follows:

$$P(X{=}k) = P(X{=}k \mid \text{state 1})*P(\text{state 1}) + P(X{=}k \mid \text{state 2})*P(\text{state 2})$$

P(X=k | state j) is determined by a simple Poisson process with average rate $\lambda_j$, $j = 1, 2$. We denoted P(state 1) by $\alpha$ in the previous section, so we can write:

$$P(X = k) \;=\; \alpha \frac{(\lambda_1 T)^k \, e^{-\lambda_1 T}}{k!} + (1 - \alpha) \frac{(\lambda_2 T)^k \, e^{-\lambda_2 T}}{k!} \qquad (4.9)$$

The above equation expresses the pdf of the counting process over a fixed observation interval T for a single 2-state MMPP minisource.

The pdf of the traffic from a superposition of N minisources can be calculated by using either the convolution of N pdf's of minisources, or the Z-transform. The Z-transfrom of (4.9) can be written as:

$$P_1(z) = \alpha\, e^{(z-1)\,\lambda_1\,T} + (1-\alpha)\, e^{(z-1)\,\lambda_2\,T} \tag{4.10}$$

Hence, the Z-transform of the pdf of the aggregated traffic from N independent minisources is:

$$P_N(z) = [\alpha\, e^{(z-1)\,\lambda_1\,T} + (1-\alpha)\, e^{(z-1)\,\lambda_2\,T}]^N \tag{4.11}$$

Denote the total number of arrivals of the aggregate traffic in an observation interval by a random variable $X_N$. By taking the inverse transform of Equation (4.11), we obtain:

$$P(X_N = k) = \sum_{i=0}^{N} \frac{N!}{i!\,(N-i)!}\, \alpha^i\, (1-\alpha)^{N-i}\, \frac{[i\,\lambda_1\,T + (N-i)\,\lambda_2\,T]^k}{k!}\, e^{-[i\,\lambda_1\,T+(N-i)\,\lambda_2\,T]} \tag{4.12}$$

By substituting $\lambda_1$ and $\lambda_2$ given by Equations (4.7) and (4.8) into Equation (4.12), we will have a formula for the pdf of the arrival rate in terms of our alternative set of parameters $[\ \overline{\lambda}\ \ d\ \ IDC(\infty)\ \ \alpha\ \ N]$.

In Figures 4.2 and 4.3, the results of the Equation (4.12) are compared to the measured pdf from simulation results, for two cases: a single minisource (N=1), and a superposition

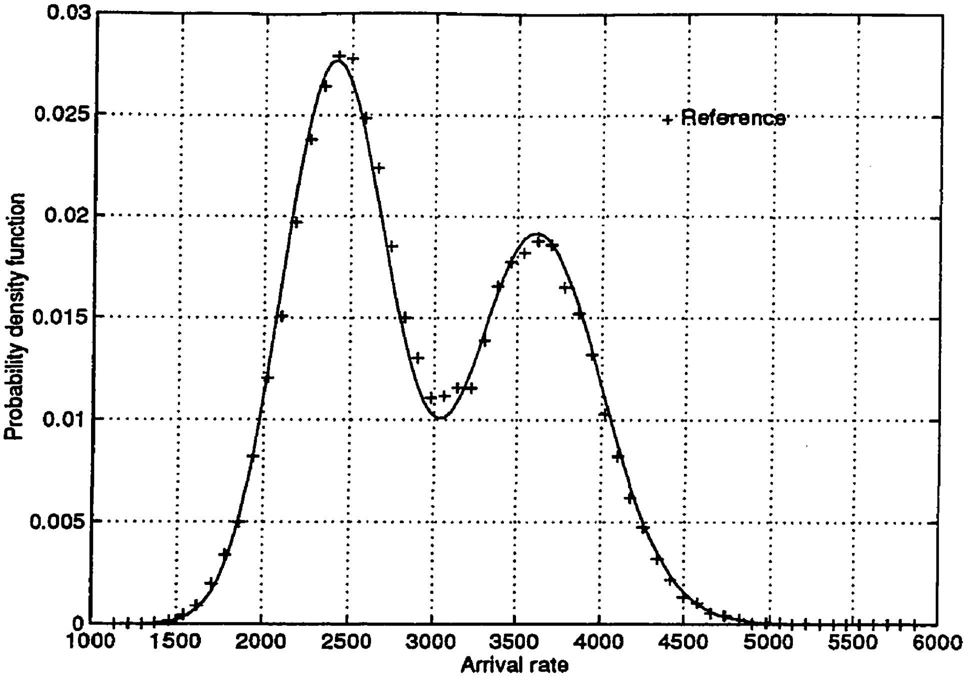

Figure 4.2: Accuracy of pdf estimation for N=1

of 6 minisources (equivalent to a 7-state MMPP). The analytical and simulation results are in a remarkably good agreement.

If the observation interval is not sufficiently small to neglect the effect of state change, Equation (4.12) becomes invalid. On the other hand, the observation interval cannot be arbitrarily reduced because of the effect of quantization error in the pdf measurement of the samples. Our simulations indicated that an observation interval equivalent to one tenth of the average sojourn time at each state provides adequate results. As an example of the cases where this approximation fail, consider a case of a traffic from a superposition of two 2-state MMPP minisources with $\lambda_1 = 2000$ cps, $\lambda_2 = 1000$ cps, $r_1 = 40 \ s^{-1}$, $r_2 = 30 \ s^{-1}$ and Frame time of 25 miliseconds. Figure 4.4 shows the pdf calculated from Equation (4.12) (continuous line) and the one determined from the simulation (dotted line). The curves show a noticable difference, due to the fact that here the observation interval is comparable

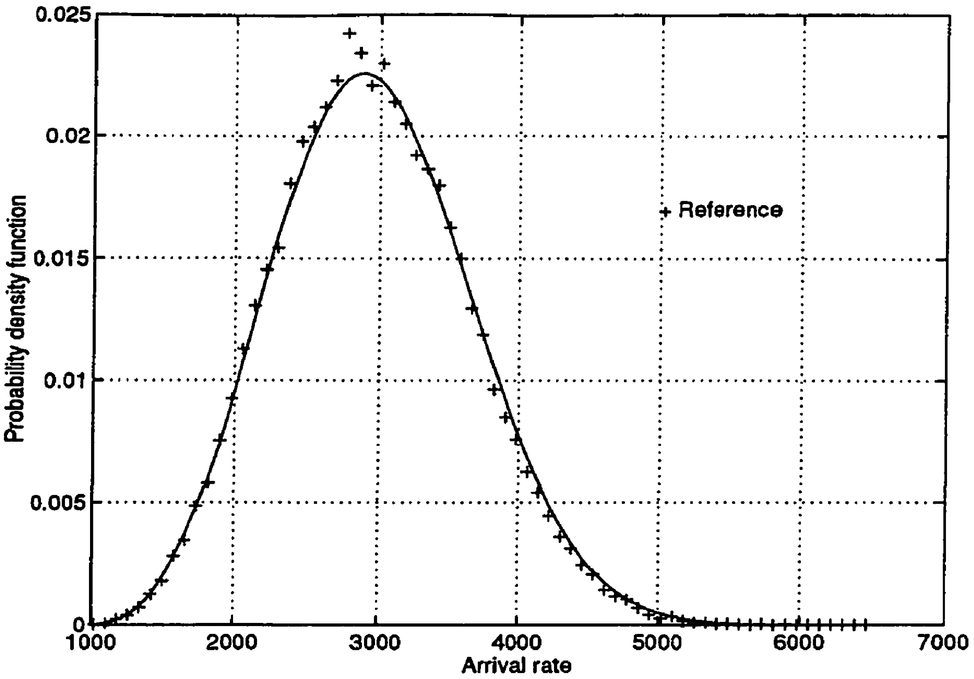

Figure 4.3: Accuracy of pdf estimation for N=6

to mean sojourn times.

It is also worth noting that in the above there is no restriction on the hidden regime of state changes. In other words, the model is just assumed to be a general switched Poisson process. Hence, the Equation (4.12) is also applicable to other switched Poisson processes such as DMPP (Deterministic Modulated Poisson Process), PMPP (Pareto Modulated Poisson Process), etc.

### 4.4.3 Parameter Estimation

We select a fixed observation interval of T called *frame time*, over which the number of arrivals is counted. Denote the number of arrivals in a frame time by a random variable $X_N$. From the sequence of measured numbers of arrivals in a frame time (i.e. measured traffic samples), we compute:

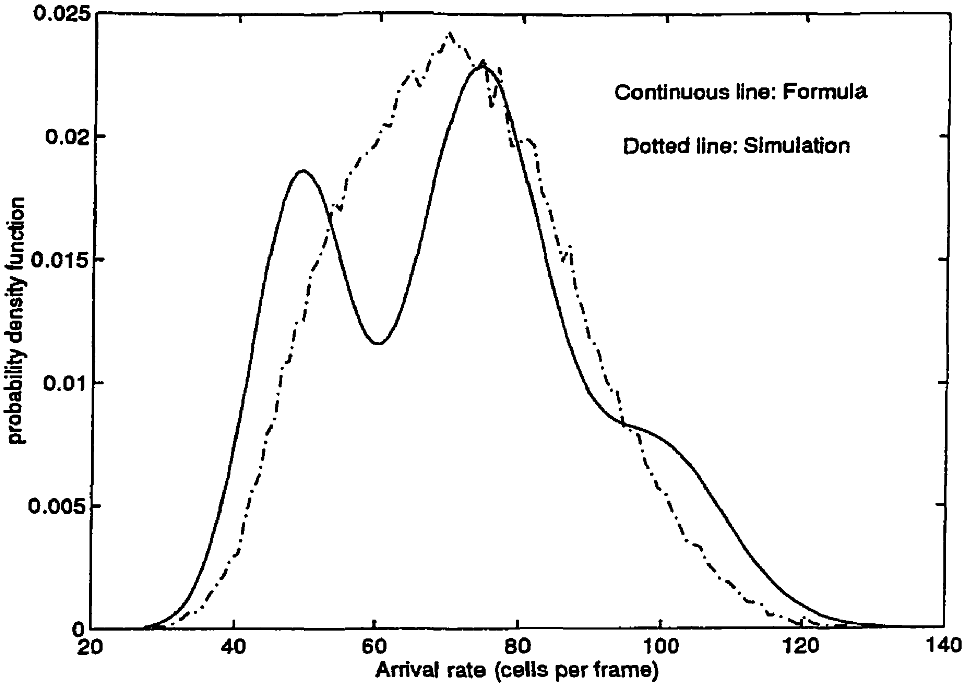

Figure 4.4: A test case where the pdf approximation fails

$$\overline{\lambda} = \frac{E[X_N]}{T} \quad IDC(t) = \frac{VAR[X_N(0:t)]}{E[X_N(0:t)]}$$

Based on the constructed IDC(t) curve, the values of $d$ and $IDC(\infty)$ can be estimated using Equation (3.2). The estimation can be done by the same maximum likelihood approach that we explained in previous chapter.

Using three derived parameters ($\overline{\lambda}$, $d$, $IDC(\infty)$ ), the remaining two parameters ($N$ and $\alpha$) can be estimated from the pdf of arrival rate, based on the minimization of the mean square error for all of the samples. To solve for N and $\alpha$ in the complex, non-linear Equation (4.12), we used an iterative optimization algorithm. The problem can have several sets of solutions for ($N$, $\alpha$) which give close results for the pdf. By increasing N from 1 and optimizing for $\alpha$ for the best match, this procedure guarantees an optimum solution

| Case | $N$ | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ (1/s) | $r_2$ (1/s) |
|---|---|---|---|---|---|
| Case 1 | | | | | |
| Reference | 2 | 2000 | 1000 | 2 | 1.5 |
| Estimated | 2 | 2024.3 | 998.3 | 2.12 | 1.535 |
| Case 2 | | | | | |
| Reference | 4 | 3000 | 500 | 2 | 1.5 |
| Estimated | 4 | 3027.8 | 500.9 | 1.96 | 1.42 |
| Case 3 | | | | | |
| Reference | 8 | 1200 | 800 | 1.8 | 1.2 |
| Estimated | 5 | 1817.9 | 1318.9 | 1.664 | 1.255 |

Table 4.2: Test cases for pdf-based matching technique

set with the lowest possible value of N and hence, the simplest model.

### 4.4.4 Illustrative Results

We have performed several simulation cases to examine the effectiveness of the introduced modeling technique. We generated cells from a superposition of MMPP mini-sources with known parameters and used the introduced modeling technique to estimate the parameters. We found the technique remarkably accurate as shown in the following table:

In Case 3, the number of minisources in the derived model is different from that of reference process. The reason is that our model captures the minimum possible number of states to model the input traffic. In multi-dimentional space formed by parameters of multi-state MMPP, the solution for matched parameters is not unique. Sometimes there are several set of parameters with the same pdf and IDC, and our technique selects the one with the lowest number of states. However, both models show the same queueing performance. Hence, from queuing point of view they are equivalent. Figures 4.5 and 4.6 show the probability of buffer overflow at a load of 0.84 and the average queue length versus traffic load, respectively, for both reference and derived models. The results are in a

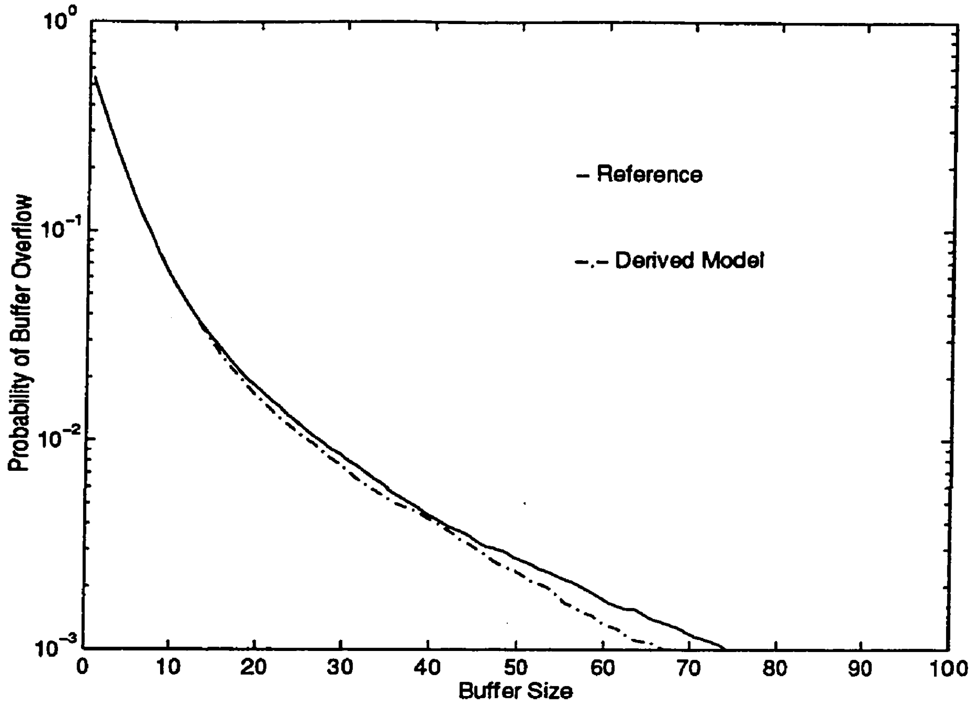

Figure 4.5: Buffer occupancy for both models

remarkably good agreement.

In Section 4.6 we will examine some case studies to show the accuracy of our pdf-based technique.

## 4.5 Approximation of the Slope of the Probability of Cell Loss for a multiple-state MMPP

In Section 3.5 we estimated the probability of loss for a 2-state MMPP/D/1 queue. Now here we extend it to calculate the slope of the probability of cell loss for our multiple-state MMPP model in a /D/1 queue. The same as Section 3.5, Equation (3.39) may be used for Laplace-Stieltjes funstion of the queue length in the *burst* region. For our model of a superposition of N homogenous 2-state MMPP with parameters $[\lambda_1, \lambda_2, r_1, r_2, N]$,

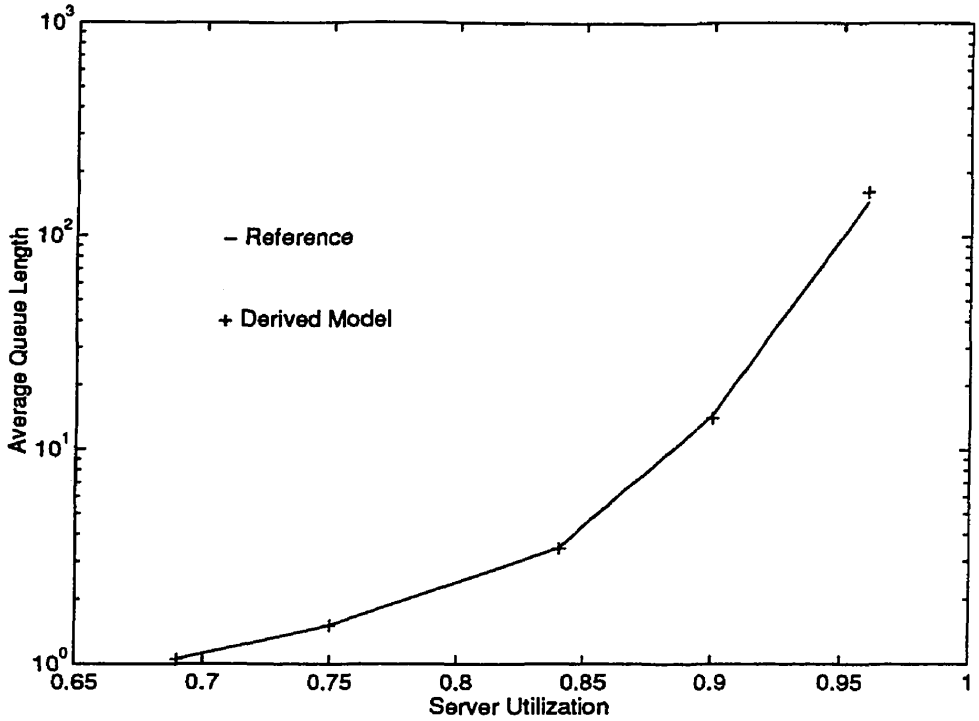

Figure 4.6: Average queue length vs. load for both models

matrices Q (transition rates) and $\Lambda$ (arrival rates) can be easily computed as follow [18]:

$$\Lambda = \mathrm{diag}(j\,\lambda_1 + (N-j)\,\lambda_2) \qquad j = 0:N \tag{4.13}$$

$$\begin{aligned} &Q(j,j) = -j\,r_1 - (N-j)\,r_2 \\ &Q(j,j+1) = (N-j)\,r_2 \qquad j = 0:N \\ &Q(j,j-1) = j\,r_1 \end{aligned} \tag{4.14}$$

Now assume that the queue length survivor function consists of summation of exponential terms. Obviously, the term having the largest negative exponential factor determines the slope of the survivor function in burst region. Therefore, the slope is the largest negative root of the denominator of D(s) (or the closest one to zero) in (3.39). The poles of D(s) can be computed by equating the determinant of the matrix $[sI + Q - \Lambda, (1 - H(s))]$ to zero.

Therefore if $s_l$ denotes the slope of probability of loss, we have:

$$|\frac{s_1}{h} I + Q - \Lambda (1 - H(\frac{s_1}{h}))| = 0 \tag{4.15}$$

For MMPP/D/1 queues, $H(s) = e^{-s\,h}$ where h denotes the cell service time. Using McLaurin series to represent $H(s) = e^{-s\,h}$, we can approximately use the first three terms, i.e., $e^{-s\,h} = 1 - h\,s + \frac{1}{2}\,(h\,s)^2$. Equation (4.15) can be re-written in an approximated form:

$$|\frac{s_1}{h} I + Q - \Lambda (s_1 + \frac{1}{2} s_1^2)| = 0 \tag{4.16}$$

The value of $s_1$ can be numerically calculated from Equation (4.16).

To assess the accuracy of the approximation, we compared the analytical and simulation results. We considered MMPP/D/1 queues. We first obtained the simulation results for the reference cases. Next, we modeled the traffic sources by the proposed multiple-state MMPP, and derived the corresponding parameters based on the traffic samples generated by the reference sources. Using the derived parameters, we subsequently approximated the slope of the probability of cell loss using Equation (4.16).

In Figures 4.7 and 4.8 two examples are shown. The traffic samples are generated by OPNET simulator from an MMPP source. The probability of loss for G/D/1 queue is obtained using simulation. The solid and dotted lines show the simulation and analytical results derived by Equation (4.16), respectively. For Case 1, the reference source is a superposition of 8 homogenous 2-state MMPP minisources with parameters $\{\lambda_1 = 1200$ cells/s, $\lambda_2 = 800$ cells/s, $r_1 = 1.8\ s^{-1}$, $r_2 = 1.2\ s^{-1}\}$. The traffic load is 0.9. The slope of survivor function calculated by our technique is $s = -0.011$. For Case 2, the reference source is a superposition of 4 homogenous 2-state MMPP minisource with parameters $\{\lambda_1 = 3000$ cells/s, $\lambda_2 = 500$ cells/s, $r_1 = 2\ s^{-1}$, $r_2 = 1.5\ s^{-1}\}$. The traffic load is 0.8. The slope of

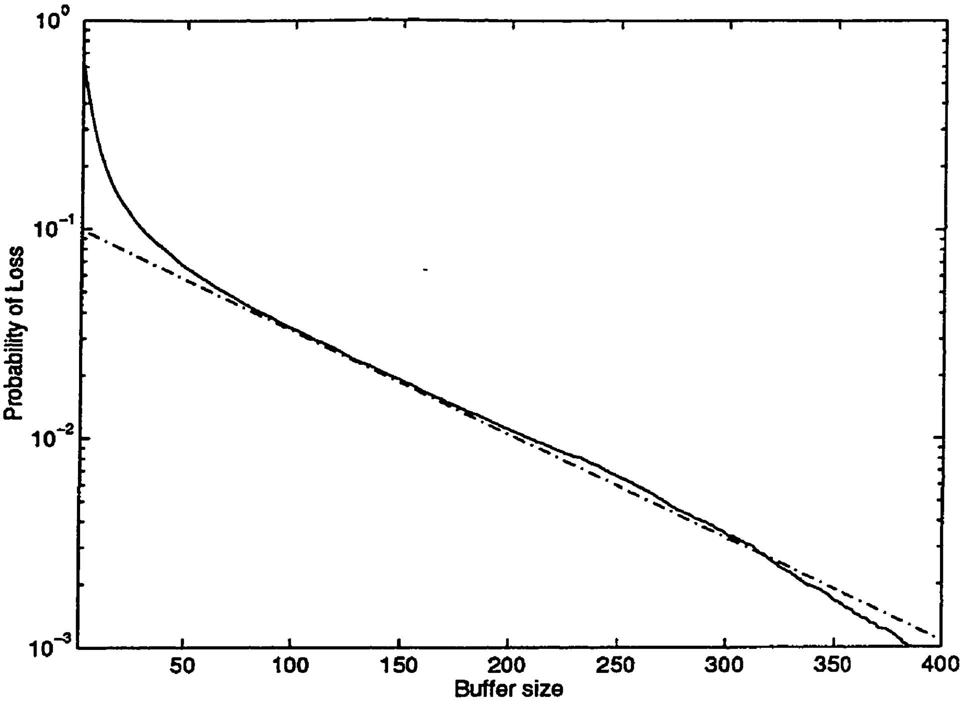

Figure 4.7: Probability of Cell Loss: Comparison of the analytical and simulation results (Test Case 1)

survivor function calculated by our technique is $s = -9 \times 10^{-4}$. Figures 4.7 and 4.8 indicate a good agreement between the analytical and simulation results.

We noticed that the approximation works fine for a model with fewer states and under heavy load. In particular, as the traffic load decreases, the accuracy of the results reduces. This effect can be explained by this fact that our approximation was valid for burst region of survivor function. When the load decreases, the queue mainly stays in cell region instead. Also when the number of states increases, due to the increased number of corresponding poles of the transfer function and the effect of nearby poles, the accuracy is also reduced.

Now let us have some examples which show the power of our technique in modeling ATM multimedia traffic.

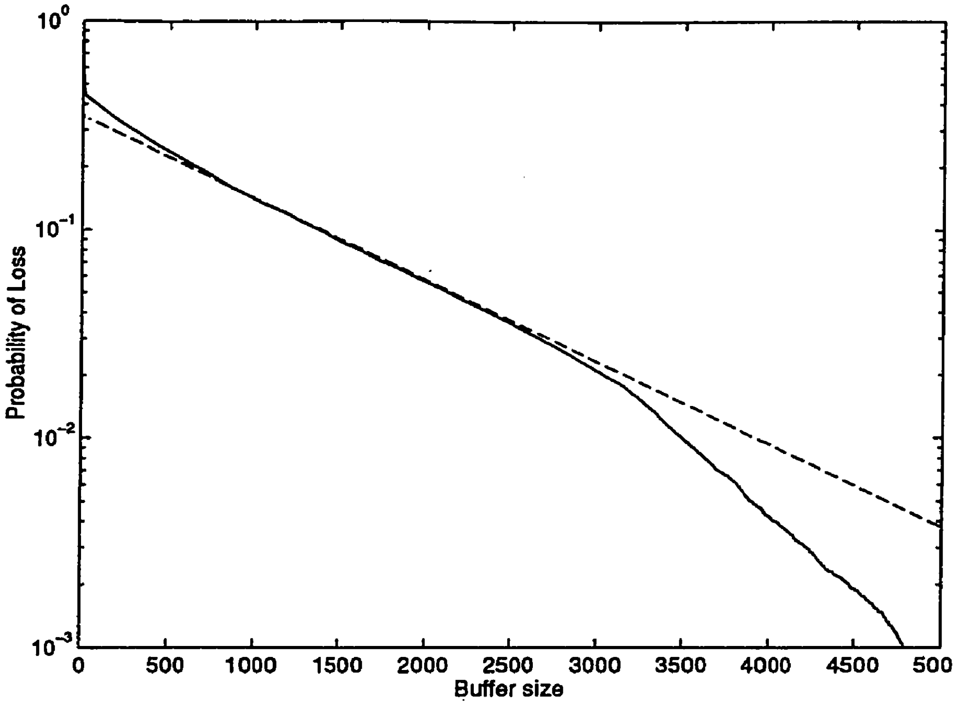

Figure 4.8: Probability of Cell Loss: Comparison of the analytical and simulation results (Test Case 2)

### 4.6 Case Studies

Here we will study two different cases and show how our technique can effectively model the traffic in ATM networks. The first case is a simple ATM multiplexer where voice, video and data traffic are mixed [28]. We try to model the aggregate traffic by a multiple-state MMPP. In the second case, the traffic inside ATM switching networks is studied [26]. For this case we build a network consists of multiplexers and switch and use simulations to show how well our model predicts the queueing performance. In the last case we get a sample of video traffic and represent it by our multiple-state MMPP model.

| Model Parameters | Reference Model |  |  | N+1-state MMPP model |
|---|---|---|---|---|
|  | Video | Voice | Data |  |
| N | 1 | 1 | 1 | 3 |
| $\lambda_1$ (cps) | 7512.4 | 7404.1 | 7612.7 | 7557.2 |
| $\lambda_2$ (cps) | 5996.9 | 6091.0 | 5925.7 | 5986.1 |
| $r_1$ ($s^{-1}$) | 1.5453 | 4.2262 | 0.2857 | 0.8745 |
| $r_2$ ($s^{-1}$) | 1.5510 | 4.3368 | 0.2857 | 0.8441 |

Table 4.3: Model parameters for case study 1: ATM Multiplexer

### 4.6.1 Case study 1: ATM multiplexer

Here we evaluated the performance of a G/D/1 queue representing a multiplexer by simulation. The input to the multiplexer is an aggregate multimedia ATM source considered to be a superposition of three components: voice, video and data. We investigated the queueing performance for two cases. In the first case (considered as the reference case), each of the traffic components is represented by a 2-state MMPP with different sets of parameters corresponding to their specific characteristics (voice, data or video). In the second case, we modeled the aggregate multimedia ATM source as a multiple-state MMPP. The parameters of this multiple-state MMPP are derived using the above mentioned procedure based on the measured traffic samples generated from the model in the first case. The parameters of the models in two cases are shown below: As Table 4.6.1 indicates, the aggregate traffic of voice, video and data has been represented by a superposition of 3 identical 2-state MMPP minisources, hence a four state MMPP model.

Figure 4.9 shows the simulation results for the two cases. A good agreement for the probability of loss in both cell and burst region is noticed.

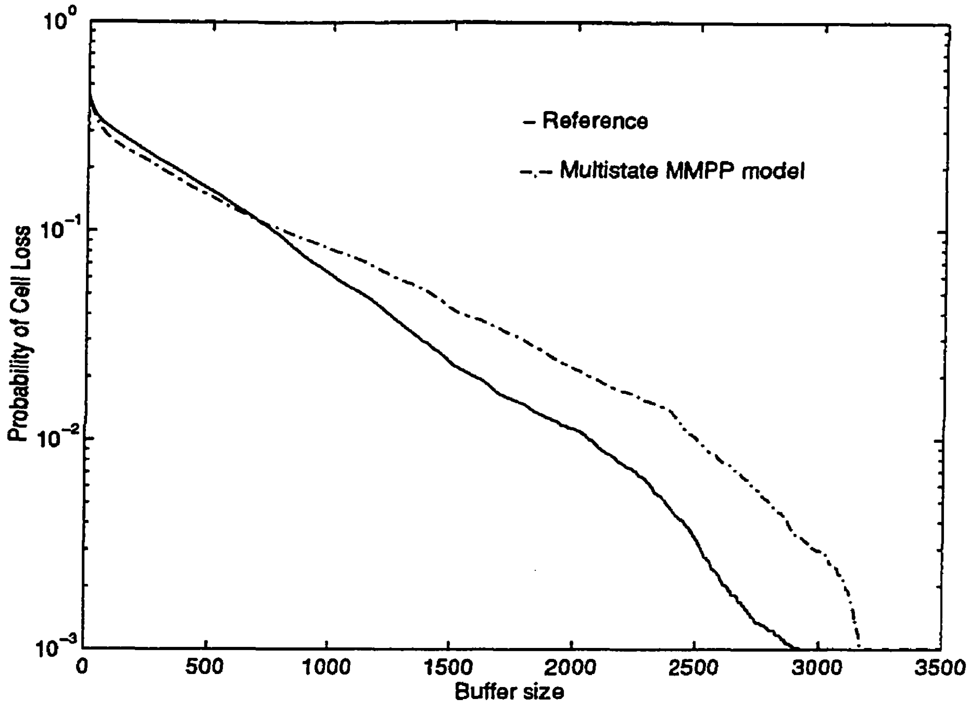

Figure 4.9: Comparing the buffer occupancy for the model and the ATM multiplexer

### 4.6.2 Case study 2: ATM switching network

Let us consider an example of multimedia ATM traffic modeling shown in Figure 4.10. Two types of real-time ATM traffic, voice and video, are multiplexed and routed through a 16X16 switch [32]. Each of the aggregate voice and video components are represented by a two-state MMPP with different parameters. All of the links are independently and identically distributed. At each multiplexer, the cells that cannot be served during a frame time will be discarded, so at the start of the next frame time the buffer is always empty. The traffic load at each multiplexer is kept around 0.63. The switch is assumed to be internally non-blocking.

The parameters of the MMPP minisource models for inputs to each switch input link

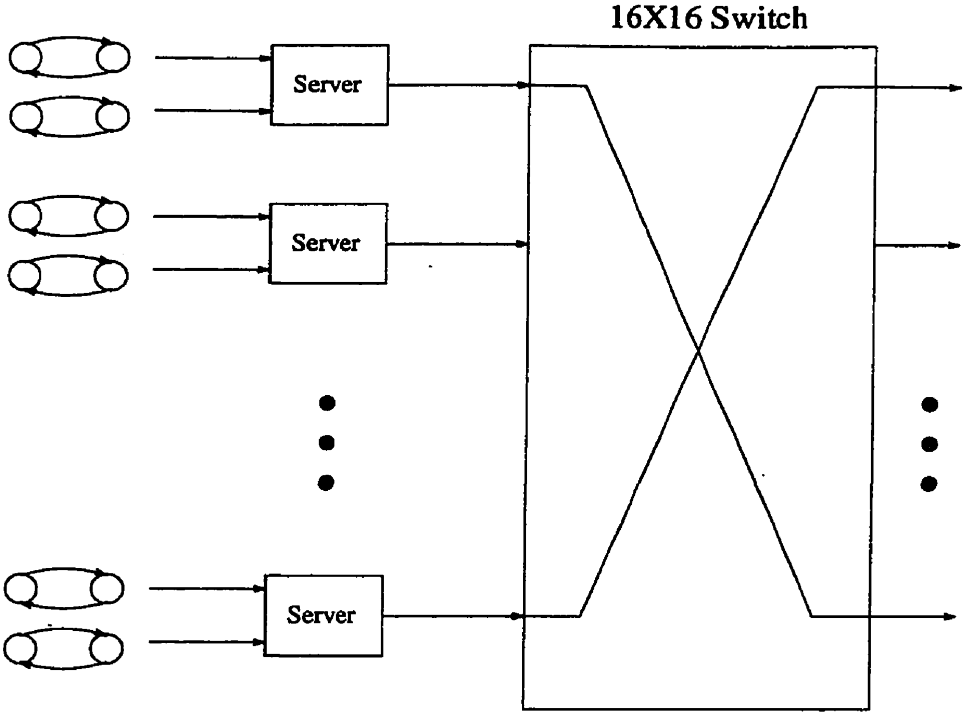

Figure 4.10: The block diagram of the system for case study #1

are listed in the table 4.6.2. One set of parameters has been used for voice and the other one for video.

We applied our technique to model the traffic at the output of each of the multiplexers, and each of the output links of the switch. From the collected simulation results on traffic samples at each point, we applied the proposed technique to obtain a model of N homogenous 2-state MMPP minisources. In the Table 4.6.2 the parameters of the dervied models for multiplexer output and switch output are shown:

Note that after the switching, due to the large number of inputs and outputs, the generated model is close to Poisson: a single 2-state MMPP with close values of mean arrival rates at each state. It is quite expected that when a large number of independent traffic streams are switched and mixed, the correlation decreases significantly and so the

| Type | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ ($s^{-1}$) | $r_2$ ($s^{-1}$) |
|---|---|---|---|---|
| Voice | 7404.1 | 6091.0 | 4.2262 | 4.3368 |
| Video | 7512.4 | 5996.9 | 1.5453 | 1.5510 |

Table 4.4: Input parameters for case study 2 : ATM switching network

| Model | N (Number of minisources) | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ ($s^{-1}$) | $r_2$ ($s^{-1}$) |
|---|---|---|---|---|---|
| Multiplexer output | 2 | 7455.2 | 6067.3 | 2.2531 | 2.1647 |
| Switch Output | 1 | 13733 | 13212 | 1.9443 | 2.5774 |

Table 4.5: Model parameters for case study 2 : ATM switching network

Poisson model is more applicable. In Figure 4.11 the IDC curves for the traffic at the multiplexer output and at the switch output have been compared. As the figure indicates, the value of IDC, which is a good indicator of interframe correlation, decreases after the switching.

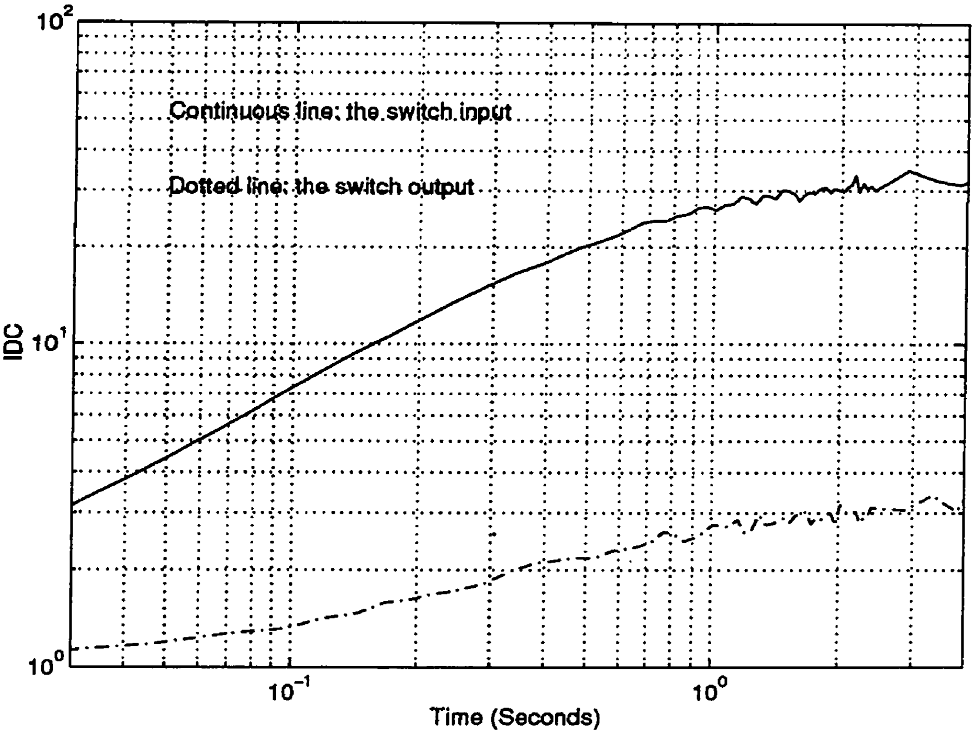

Figure 4.11: The IDC curve for the traffic at the input and the output of the switch in the case #2

Now, by using the obtained models in a separate G/D/1 queue, we performed the performance evaluation and compared the results to those of the reference model in Figure 4.10. The performance comparison is shown in Figures 4.12 and 4.13 for the outputs of the input multiplexer and the switch, respectively. As the figures indicate, a good agreement is noticed.

However, one must note that in some cases, especially when the correlation is still too

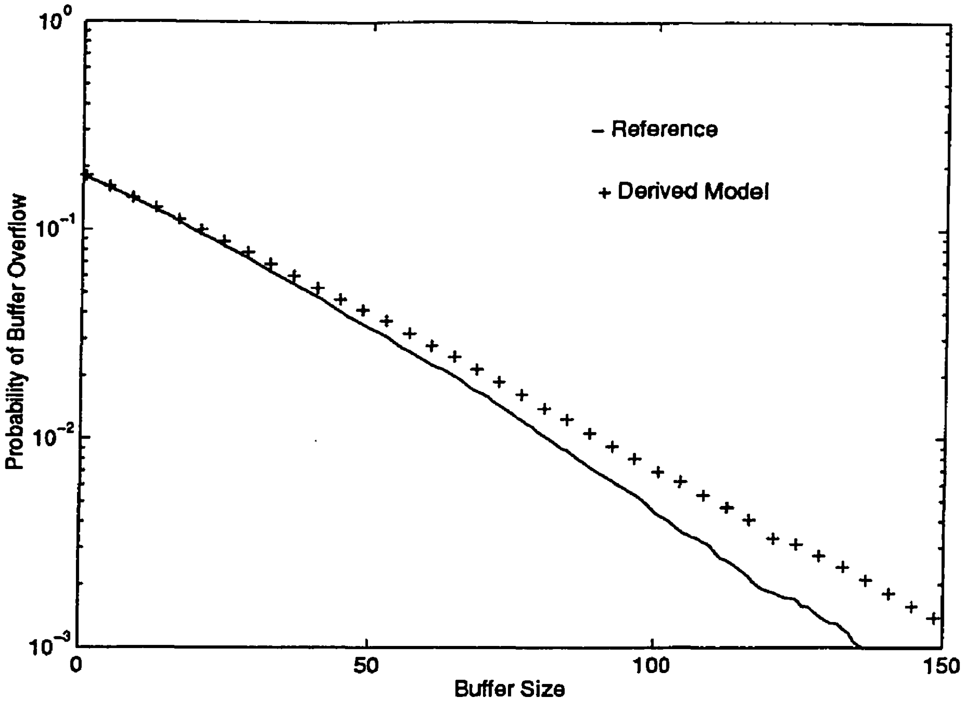

Figure 4.12: Comparing the buffer occupancy for the model and the output of the multiplexer

high after the switching, our simple model cannot represent the traffic in an efficient way and a more complicated multiple-state MMPP which does not comply to our simplifying assumption of the superposition of 2-state MMPP minisources, might be used.

### 4.6.3 Case study 3: Video VBR traffic

s a final case, here we are going to study the performance of the model to represent Video VBR traffic.

We have two streams of video VBR traffic which we call STRM#1 and STRM#2. The first trace, STRM#1, is a soccer game and STRM#2 is a movie. The files are available for public use on *ftp://www-info3.informatik.uni-wuerzburg.de/pub/MPEG/*. In Table 4.6.3 you can find the technical specifications of the streams [37].

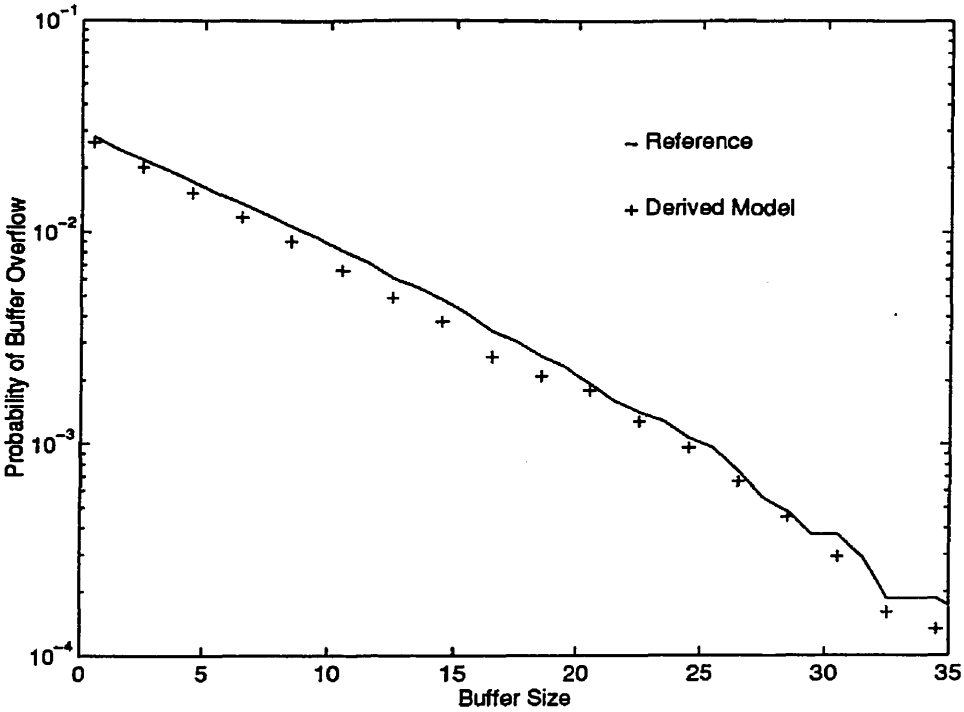

Figure 4.13: Comparing the buffer occupancy for the model and the output of the switch

Each trace contains a total number of 40000 frame which is equivalent of a time length of a bit more than half an hour.

We took the whole stream including I, B and P frames and applied our technique to model it with a multiple-state MMPP. In Table 4.6.3 the parameters of the model derived by Pdf-based matching for each of the streams are shown.

In the above table, $N$ denotes the number of minisources and $\lambda_1$, $\lambda_2$, $r_1$ and $r_2$ denote the parameters of the 2-state MMPP minisource. The model will be a superposition of $N$ homogeneous 2-state MMPP minisources with the above parameters, thus an $N + 1$-state MMPP.

Now first let study the main characteristics of each of the streams and the corresponding models. In Figures 4.14 and 4.15 the IDC curves of the model and the video stream for

| | |
|---|---|
| Coding Scheme | MPEG-1: (Berkeley MPEG-encoder version 1.3) |
| Capture rate: | 25 frame per seconds |
| Encoder Input: | 384 x 288 pel |
| Color format: | YUV (4:1:1, resolution of 8 bits) |
| Quantization values: | I=10, P=14, B=18 |
| Pattern: | IBBPBBPBBPBB |
| GOP size: | 12 |
| Motion vector search: | 'Logarithmic' / 'Simple' |
| Reference frame: | 'Original' |
| Slices: | 1 |
| Vector/Range: | half pel / 10 |
| Total number of frames in trace: | 40000 |

Table 4.6: The technical specifications of the video traces

| Trace | Model parameters | | | | |
|---|---|---|---|---|---|
| | $N$ | $\lambda_1$ (cps) | $\lambda_2$ (cps) | $r_1$ ($s^{-1}$) | $r_2$ ($s^{-1}$) |
| STRM#1 | 8 | 845.2356 | 67.1195 | 0.3561 | 0.0782 |
| STRM#2 | 3 | 344.6491 | 76.5825 | 0.2088 | 0.3131 |

Table 4.7: Model parameters for video traces

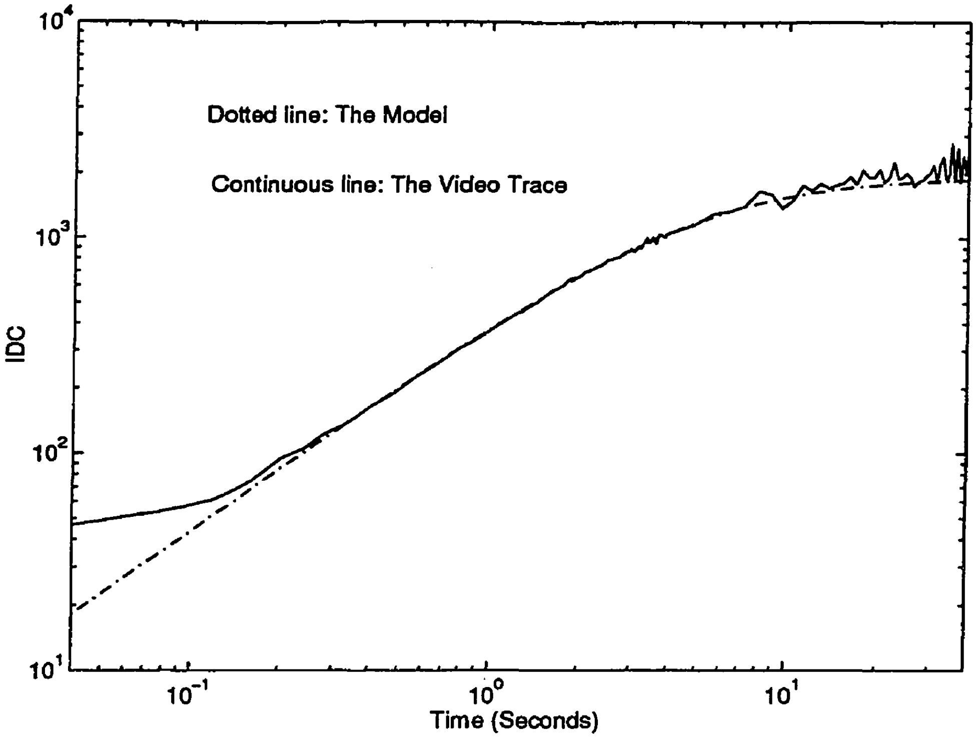

Figure 4.14: IDC curves for the video stream STRM#1 and its MMPP model

each trace have been shown. A very good agreement is noticed, expectedly. Although, the value of variance at small lags are different. Apparantly, the video traces have specific IDC characteristics at small time lags, like a small fall until a minimum value before a monotonic rise until saturation, which is uncapturable by the MMPP model. The minimum value of IDC for MMPP is located at $t = 0$ and is equal to 1.

In Figures 4.16 and 4.17 the probability density function of the arrival rate for the model and the video stream for each trace have been shown. Although the detailed shape of the curves look a bit different but it successfully catches the regions in which a high probability exists. It looks possible to have a general multiple-state MMPP which can have exactly the same pdf as the video traces have, however, our simplified model was unable to approach closer to the reference pdf.

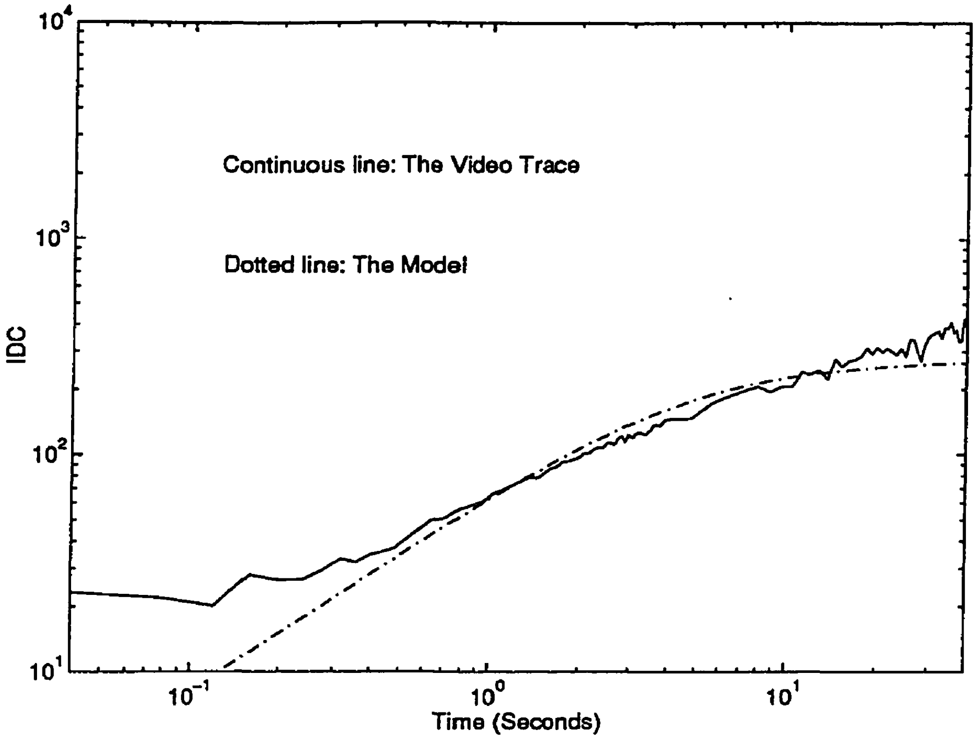

Figure 4.15: IDC curves for the video stream STRM#2 and its MMPP model

Now we used each trace and its computed model in a separate G/D/1 queue to compare their queueing performance. In Figure 4.18 you observe the curves of the mean queue length versus time for video stream STRM#1 under various traffic loads. The results have been obtained by using simulation. In the figure, the dotted line indicates the performance of our model, and the continuous line indicates the performance of the real video trace SRTM#1 in separate G/D/1 systems. The values of mean queue lengths for both systems gradually converge to the same value. In high traffic load, the system is not still in steady-state situation because the length of the video trace was limited.

The same curves can be observed for STRM#2 in Figure 4.19. Here also the effect of transient behaviour can be noticed. The results for the first video stream is closer which shows our model was more applicable in that case.

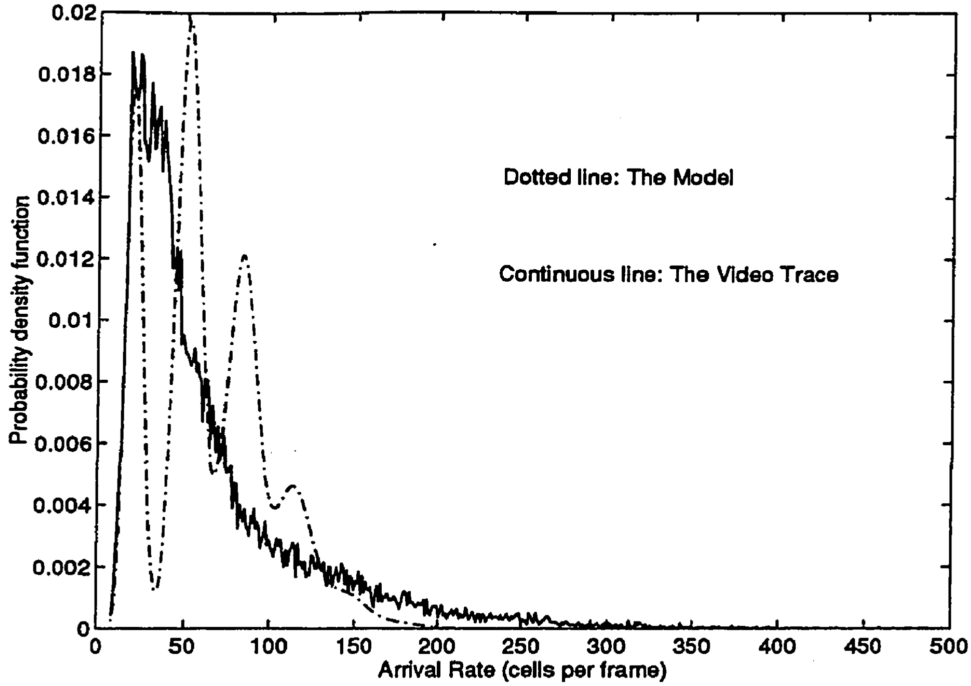

Figure 4.16: Pdf of the video stream STRM#1 and its MMPP model

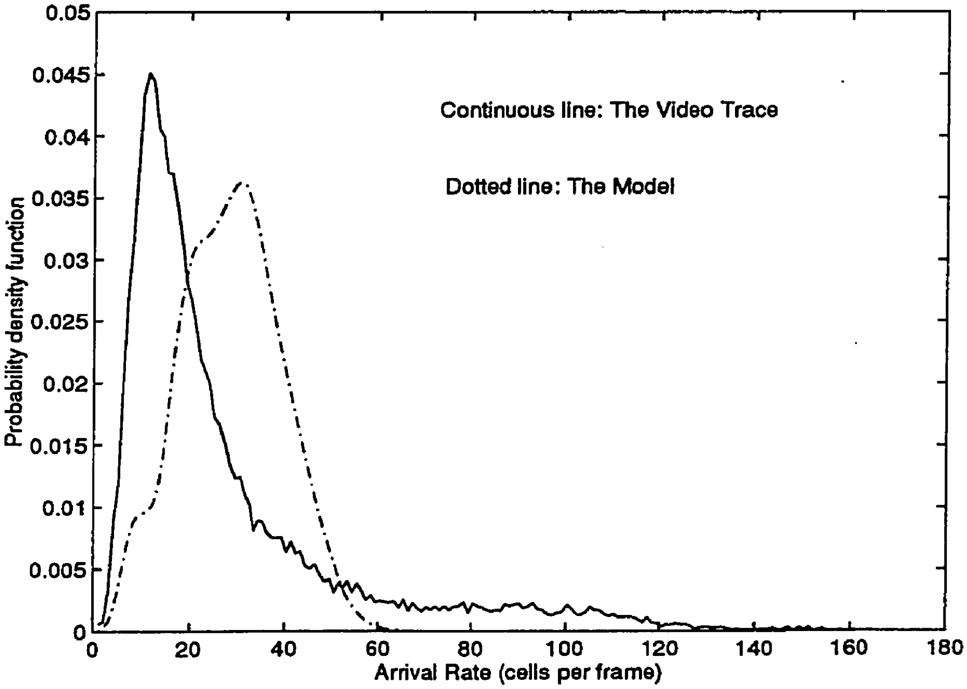

Figure 4.17: Pdf of the video stream STRM#2 and its MMPP model

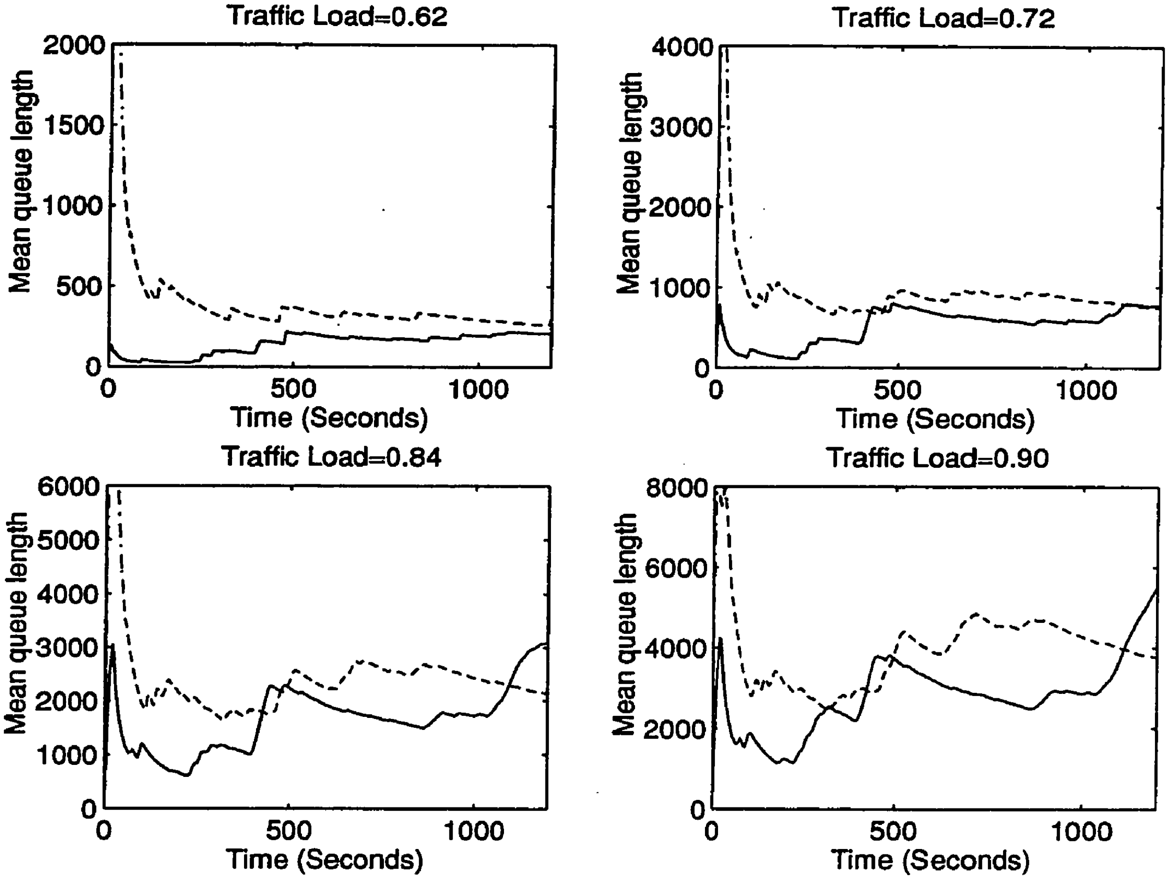

Figure 4.18: Queueing performance of the video stream STRM#1 and its corresponding model

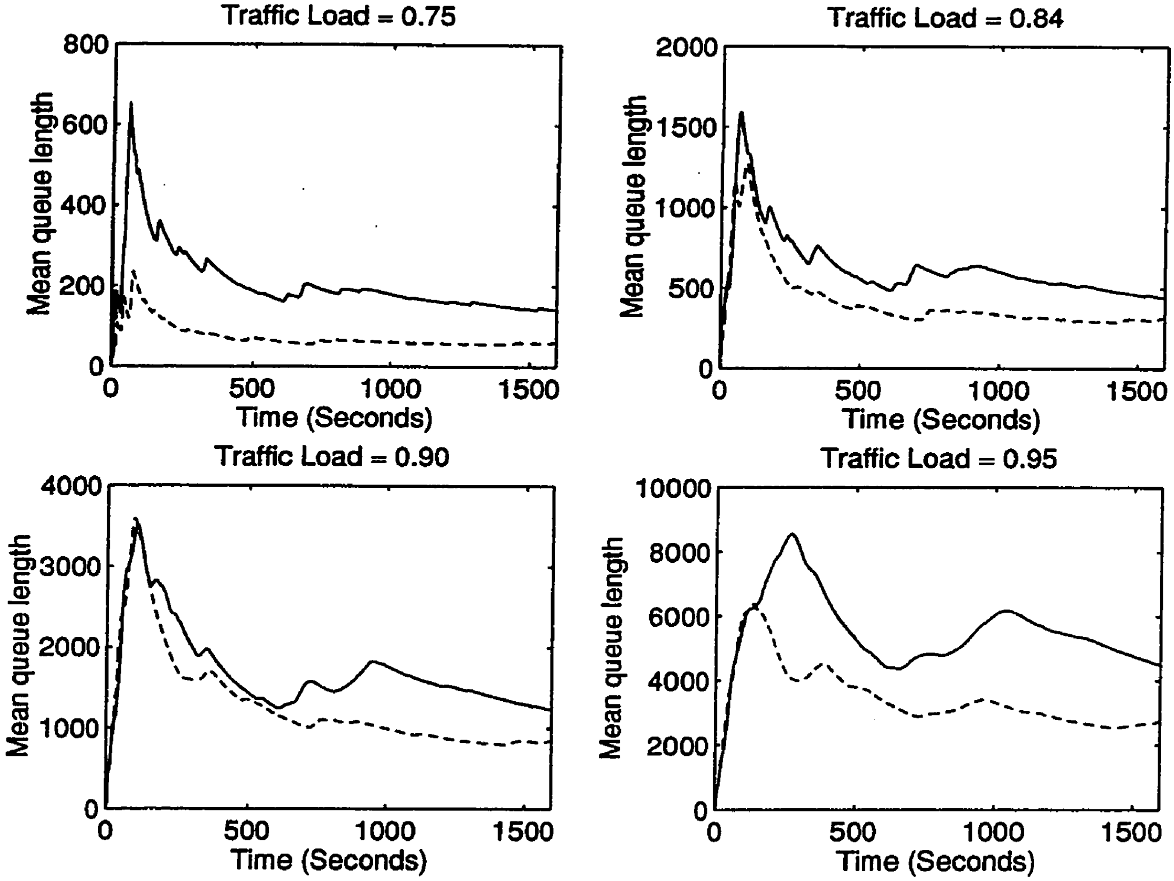

Figure 4.19: Queueing performance of the video stream STRM#2 and its corresponding model

Due to the fact that in both cases the simulation could not be continued to the steady-state situation, comparing the curves of probability of loss is not possible. Therefore we just used the value of mean queue length as our performance indicator here.

# CHAPTER 5
# Conclusion and Future Works

In this thesis we studied the performance of Markov-Modulated Poisson Process (MMPP) to represent the multimedia ATM traffic. First we applied the 2-state MMPP model for aggregated voice traffic and compared various matching techniques for deriving the parameters of the MMPP model in order for the model to be able to predict the queueing performance of the aggregated traffic. We observed that a simple overload-underload IDC matching technqiue provides us with a satsifactory accurate prediction of the performance without need to go for lengthy, complicated, time and computing power consuming techniques that use inverse laplace transforms.

In the next step, we generalized a moment-based technique to make it capable of matching a 2-state MMPP model to a general, arbitrary ATM traffic. By this, we were able to take a sequence of traffic samples and model it by a 2-state MMPP. Furthermore, an approximation for the probability of loss in 2-state MMPP/D/1 queue was also derived to avoid the lengthy and complicated Matrix Geometric techniques. At this point, we are able to predict the queueing performance of a given traffic stream, provided that we have enough number of samples and the 2-state MMPP model is applicable.

We found that there are certain cases were two states are not enough to represent the changes in the arrival rate. Therefore we studied the multiple-state MMPP case. In order to overcome the complicated problem of parameter-fitting for a general multiple-state case, we introduced a special case, a superposition of N 2-state MMPP minisources, which is an equivalent of a special N+1-state MMPP. We presented a pdf-based technique to derive the parameters of this model from the traffic samples. We showed that IDC curve and pdf of arrival rate are enough to derive the parameters. We completed our job by suggesting a technique to estimate the slope of the curve of the probability of loss for this special multiple-state MMPP/D/1 queue. Several case studies were also presented to show the power of the technique.

There are a number of areas in which this work can be continued. While an optimization algorithm for pdf-based matching was offered, still the parameters of the N+1-state MMPP model cannot be explicitly expressed in terms of the values of pdf and IDC. One may try to come up with a modification in the lengthy fitting procedure in order to simplify the process of finding the model parameters. Particularly in finding the number of states and steady-state probability of staying in state 1 from probability density function of arrival rate, some approximations may help.

We found that while the model performs satisfactorily in most cases, it seems that some more complicated models like a general multiple-state MMPP may be needed in certain cases. Therefore one way to continue this work, is to generalized the model and to try to find a fitting technqiue for general multiple-state case. Especially in the case of video, some modifications may help us to have a more accurate model to represent VBR video traffic.

Finally, the approximation for pdf which is used in this technique is applicable to any other switched Poisson process like DMPP and PMPP too. If the IDC curves of these models are known, one may want to try to find a fitting technqiue for these models. In particular, PMPP which is capable of capturing long range dependency looks an attractive model to study. Due to the expansion of Internet and the need to come up with a model capable of capturing self similarity and long range dependency, a PMPP matching technique for deriving the parameters of the model from the traffic samples will be an interesting area for research.

# Bibliography

[1] Ibrahim W. Habib and Tarek N. Saadawi, "Multimedia Traffic Characteristics in Broadband Networks", *IEEE Communication Magazine*, pp. 48-54, July 1992.

[2] Jaime Jungok Bae and Tatsuya Suda, "Survey of Traffic Control Schemes and Protocols in ATM Networks", *Procedings of IEEE*, Vol. 79, No.2, pp. 170-189, February 1991.

[3] G. D. Stamoulis, M. E. Anagnostou and A. D. Georgantas, "Traffic source models for ATM networks: a survey", *Computer Communications*, Vol. 17, No. 6, pp. 428-438, June 1994.

[4] Victor S. Frost and Benjamin Melamed, "Traffic Modeling for Telecommunications Networks", *IEEE Communication Magazine*, pp. 70-81, March 1994.

[5] John P. Cosmos et al, "A Review of Voice, Data and Video traffic Models for ATM", *European Transactions on Telecommunication, Special Issue on Teletraffic Research for B-ISDN in the RACE program*, Vol. 5, No. 2, pp. 139-154, March-April 1994.

[6] Selvakumaran Subramanian, Tho Le-Ngoc, "Traffic modeling in a multimedia environment", *Proc. of Canadian Conference on Electrical and Computer Engineering (CCECE'95)*, Montreal, pp. 838-841, September 5-8, 1995.

[7] Selvakumaran N. Subramanian, "Traffic Modeling in a multimedia environment", M.A.Sc Thesis, Concordia University, 1996.

[8] W. E. Leland, S. M. Taqqu, W. Willinger and D. V. Willson, "On the self-similar nature of Ethernet Traffic", *IEEE/ACM trans. on Networking*, Vol.2, No.1, pp. 1-14, February 1994.

[9] D. R. Cox, P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Matheun, London, 1966.

[10] Mischa Schwartz, *Broadband Integrated Networks*, Prentice-Hall, 1996.

[11] Leonard Kleinrock, *Queueing Systems*, John Wiley and Sons, 1975.

[12] D. Le Gall, "MPEG: A Video Compression Standard for Multimedia Applications", *Communications of the ACM*, Vol. 34, pp. 47-58, April 1991.

[13] Daniel P. Heyman, T. V. Lakshman, "Source Models for VBR Broadcast-Video Traffic", *IEEE/ACM Transaction on Networking*, Vol. 4, No. 1, pp. 40-48, February 1996.

[14] Basil Maglaris, Dimitris Anastassiou, Prodip Sen, Gunnar Karlson and John Robbins, "Performance Models of Statistical Multiplexing in Packet Video Communications", *IEEE Tran. on Communication*, Vol. 36, No. 7, pp. 834-844, July 1988.

[15] Reto Grunenfelder, John P. Cosmos, Sam Manthorpe and Augustine Odinma-Okafor, "Characterization of Video Codecs as Autoregressive Moving Average Processes and Related Queueing System Performance", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 3, pp. 284-293, April 1991.

[16] Daniel Reininger, Benjamin Melamed and Dipankar Raychaudhuri, "Variable Bit Rate MPEG Video Characteristics, Modeling and Multiplexing", *ITC 14* (Editors: J. Labetoulle and J. W. Roberts), Elsevier Science B.V., 1994.

[17] H. Heffes and D. M. Lucantoni, "A Markov Modulated characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance", *IEEE J. Select. Areas Commun.*, Vol. SAC-4, No. 6, pp. 856-868, September 1986.

[18] W. Fischer, K. Meier-Hellestern, "The MMPP cookbook", *Performance Evaluation*, 18, pp. 149-171, 1992.

[19] K.S. Meier-Hellstern, "A fitting algorithm for Markov-modulated Poisson processes having two arrival rates", *European Journal of Operational Research*, 29, pp. 370-377, 1987.

[20] Tobias Ryden, "Parameter estimation for Markov modulated Poisson processes", *Commun.Statist.-Stochastic models*, 10(4), pp. 795-829, 1994.

[21] P. Skelly, M. Schwartz, S. Dixit, "A Histogram-based model for video traffic behaviour in an ATM multiplexer", *IEEE/ACM Transaction on Networking*, Vol.1, Number 4, pp. 446-459, August 1993.

[22] J. Huang, T. Le-Ngoc, J. F. Hayes, "Broadband SATCOM system for multimedia services", *Proc. ICC'96*, Dallas, pp.906-909, 1996.

[23] A. Baiocchi, N. Blefari Melazzi, M. Listanti, A. Roveri, R. Winkler, "Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed ON-OFF Sources", *IEEE J. Select. Areas Commun.*, Vol. 9, No. 3, pp. 388-393, April 1991.

[24] S.H. Kang and D.K. Sung, "Two-state MMPP modeling of ATM superposed traffic based on the characterization of correlated interarrival times", *Proc. IEEE GLOBECOM'95*, pp.1422-1426, Nov. 1995.

[25] R. Gusella, "Characterizing the variability of arrival processes with indexes of dispersions", *IEEE J. Select. Areas Commun.*, SAC-9, pp.203-211, 1991.

[26] Shahram Shah-Heydari, Tho Le-Ngoc, "Pdf-Based Modeling of ATM Traffic by Multistate MMPP", *Proc. 19th biennual sympusium on Communication*, Kingston, Canada, May 31-June 3, 1998.

[27] Shahram Shah-Heydari, Tho Le-Ngoc, "MMPP modeling of Aggregated ATM traffic", *Proc. Canadian Conference on Electrical and Computer Engineering (CCECE'98)*, Waterloo, Canada, May 24-28, 1998.

[28] Shahram Shah-Heydari, Tho Le-Ngoc, "Multiple-State MMPP Models for Multimedia ATM Traffic", *Proc. International Conference on Telecommunications (ICT'98)*, Chalkidiki, Greece, June 22-25, 1998.

[29] S. Kim, M. Lee, M. Kim, "$\sum$-Matching Technique for MMPP Modeling of Hetrogeneous ON-OFF Sources", *Proc. Globecom '94*, pp.1090-1094, 1994.

[30] K. Sriram, W. Whitt, "Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data", *IEEE J. Select. Areas Commun.*, Vol. SAC-4, No. 6, pp. 833-846, September 1986.

[31] Sang H. Kang, Changhwan Oh, Dan K. Sung, "A Traffic Measurement-Based Modeling of Superposed ATM Cell Streams", *IEICE Trans. Commun.*, Vol. E80-B, No.3, pp. 434-441, March 1997.

[32] Tho Le-Ngoc, Tien Hy Bui, Mohamed Hachicha, "Performance of a Knockout switch for multimedia satellite communication", *Proc. IEEE GLOBECOM'97*, Phoenix, AZ, USA, November 4-8, 1997.

[33] Stephan Robert and Jean-Yves Le Boudec, "Can self-similar traffic be modeled by Markovian processes?", *COST242 Technical Document* , TD95-26, presented at the Stockholm Management Committee Meeting, May 10-11, 1995.

[34] *OPNET Manuals*, MIL3 Inc. (*http://www.mil3.com*), Washington DC, 1994.

[35] Allan T. Andersen and Bo Friis Nielsen, "An Application of Superpositions of two state Markovian Sources to the Modelling of Self-similar Behaviour", *IEEE Infocom '97 - 16th Conference on Computer Communications*, Kobe, Japan, April 7-11, 1997.

[36] G. Lindgern and U. Holst, "Recursive Estimation of Parameters in Markov-Modulated Poisson Processes", *IEEE Transaction on Communication*, Vol.43, NO.11, pp. 2812-2819, November 1995.

[37] Oliver Rose. "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems",*Proc. of the 20th Annual Conference on Local Computer Networks* Minneapolis, MN, 1995, pp. 397-406.