Capturing modes in bursty network traffic

by

Ramakrishnan Krishnaswamy

Submitted to the Faculty of the University of Kansas

in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

August 10, 2004

Abstract

Modern internet traffic is bursty in nature owing to the presence of long-range dependencies in the traffic arrival stream. Unpredictable packet losses occur at network multiplexing points due to burstiness, calling for traffic models which provide insight into the nature of traffic arrival mechanism, and that can predict the queueing behavior accurately. We propose that a queueing system with bursty input traffic can be effectively represented by a weakly stable multi-modal nearly completely decomposable continuous parameter Markov chain. We develop techniques to extract modes from traffic traces of known reliability, and study the properties of the extracted modes. We then model the multi-modal lossy queueing system using linear algebraic queueing theory techniques based on the observed properties of modes. We derive analytically tractable solutions to find the steady state vector of the system in a manner that provides an insight into the impact of different modal components on the system performance. We then verify the analytical model and the solution techniques developed by comparing the packet loss estimates obtained from the model with those obtained from the trace driven simulations employing the original traffic traces.

Dedicated to my Parents

Acknowledgments

I would like to express my sincere gratitude to my committe co-chair Dr. Victor Wallace for providing me with the oppurtunity to work in this project. I thank him for his kind support and encouragement which enabled me to complete this thesis successfully. I would like to acknowledge my committee co-chair Dr. David Petr for his valuable ideas and suggestions. I thank Dr. John Gauch for accepting to serve in my committee. I would like to specially acknowledge Dr. Appie van de Liefvoort and Dr. Kenneth Mitchell, the School of Computing and Engineering, University of Missouri, Kansas city, for educating me on the fundamentals of linear algebraic queueing theory.

*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Traffic flowing through the telecommunication networks in the pre-internet age was predominantly 'voice'. The number of calls arriving at a station, namely the counting process, approximated a Poisson or renewal process. In either case arrivals were memoryless in the Poisson case, or memoryless at renewal points, and interarrival intervals were exponentially distributed. The Poisson arrival model and exponentially distributed holding time model allowed analytically and computationally simple 'birth and death' Markov chains to be used for much of the telephone traffic modeling. An M/M/1/K chain can be used to accurately model a single server finite queue system with exponential service and Poisson arrivals yielding closed form solutions for queue length distribution, waiting time distribution, blocking probability etc.

But we don't have this comfort while modeling modern Ethernet traffic or

Internet traffic, which behave very differently from such simple Markovian models. Traffic measurements made at the Local Area Networks (LAN) and Wide Area Networks(WAN) of Bellcore's Morristown laboratories suggest that traffic exhibits variability (traditionally called 'burstiness') over multiple time scales [30]. The second order properties of the counting process of the observed traffic displayed behavior that is associated with self-similarity, multi-fractals and/or *long range dependence*(LRD). This indicates that there is a certain level of dependence in the arrival process. Near-range and long-range dependencies often manifest themselves in a network by causing frequent and irremediable packet losses and other serious effects in the network.

Dependencies and burstiness in traffic hence brought in an enormous amount of attention from researchers. They attempted to develop mathematically-based models that would help explain the nature of the systems exhibiting such phenomena and provide critical insight into the actual mechanisms that led to this behavior. Models like fractional Brownian motion, chaotic maps etc. were suited to capture the second order self-similar behavior of traffic [5, 9, 23]. Their results were difficult to get and harder to apply, and such models did not provide insight into the actual mechanism of traffic generation. Many analytically simpler modeling attempts to capture the first and second order properties of counts did not predict the queueing behavior well enough. In the late 90s researchers discussed the impact of other properties of the self-similar process, such as marginal distributions, in accurately predicting the queueing behavior. A simpler, more ac-

curate and analytically tractable model that provides more physical insight into why they are meaningful on physical grounds would help the network designers produce more effective and efficient designs.

Some ideas generated in the last decade offer promise towards crafting the model. Anderson and Nielsen [1] illustrated that continuous parameter Markov chains (cpMc) can model the dependencies in network traffic over multiple time scales; the advantage of such models is the availability of ready-made tools for analysis. Their model matched the second order properties of the self-similar process closely, but it was not sufficient for accurate prediction of queueing behavior. Grossglauser and Bolot [9] discussed both the importance of limiting the view to the finite range of time scales of interest, and the influence of marginal distributions in performance evaluation and prediction problems. From the above discussions, one can infer that both the second order and marginal properties of the process need to be matched for more accurate results. Salvador et al. [27] achieved some degree of success by using a fitting procedure that matched both the marginal distribution and autocovariance of the counting process, but a solution form that provides deep insight into the system was still missing.

Jelenkovich [11] found that in MPEG traffic there are some unstable modes (he refers to them as regimes) having conditional mean arrival rate greater than that of server capacity, even though the overall system may be stable. He calls this behavior weak stability. In general the unstable mode can be identified with a class of states in the traffic model which is rarely entered but leads to burstiness.

Hence it is possible that most of the dependent bursty Ethernet and Internet traffic streams also can be characterized by a multi-modal cpMc. Juliano [12] showed that, in the presence of highly correlated arrivals, the systems under study could be divided into subsystems such that the interaction between certain subsystems is much slower than the interactions within the subsystems. This property in Markov chains is called near complete decomposability (NCD) [2, 33]. The multi-modal systems can be NCD as there are weak inter-modal leakages between modes. Also he showed that computations on Markov chains involving NCD matrices could be simple and economical. The solution method he gave provided critical insight.

These ideas suggest that bursty behavior can be realistically and effectively captured by a weakly stable NCD cpMc. To restate, in a weakly stable NCD system the intra-modal interactions are much greater than the coupling between the modes and the arrival rate of at least one of the modes exceeds the service rate resulting in buffer overflows. This thesis explains the extraction of modes from data of known reliability (like the Bellcore traffic trace data) and then attempts to evaluate the effectiveness of the resulting model.

Organization of Thesis

Chapter 2 will discuss the concept of *long range dependence* in network traffic and the causes. The different approaches of traffic modeling will be discussed, stating the advantages of Markov models over other models. The Chapter also

covers linear algebraic queueing theory techniques used in building the analytical model. Chapter 3 will introduce the concept of modes and discuss the different mode extraction procedures highlighting their advantages and disadvantages. The properties of different modes obtained will also be discussed. Chapter 4 focusses on the step-by-step development of the analytical model and discusses the methods to obtain the solution for the model. Then the performance of the multi-modal traffic model will be studied and compared with other published models. Chapter 5 will discuss the conclusions and future work.

Chapter 2

Background

2.1   Introduction

The purpose of this chapter is to explain the nature of the dependent network traffic and the parameters and terms used in the literature to describe such traffic. Various mathematical models, their advantages and disadvantages, and supportive points in favor of Markov models will be discussed with appropriate evidences. Sections also describe the property of near-complete decomposability and the matrix exponential distributions.

2.2 Self-similarity

2.2.1 Overview and causes

Self-similarity and fractals are notions pioneered by Mandelbrot [15]. Self-similarity exists when a certain property of an object is preserved with respect to scaling in time and/or space. If an object is self-similar, the parts which form the object, when magnified, resemble the shape of the whole object [30]. Self-similar processes are very different from the Poisson or even general renewal processes. At every time scale ranging from milliseconds to perhaps several hours similar looking bursts can be observed. On the other hand in Poisson or renewal traffic, bursts tend to smooth out as the time scale is increased. The area of traffic measurements and analysis has been tremendously active since the well-known Bellcore traffic measurements, supporting the view that network traffic has self-similar scaling behavior over a wide range of time scales. This property is robust in the sense that though networks have evolved since the Bellcore measurements in terms of topology, speed and traffic composition, observed traffic has behavior consistent with self-similarity.

Several authors attribute this traffic invariant property to the TCP and HTTP network traffic. Crovella [4] shows that network traffic that is a subset of WWW transfers can show characteristics that are consistent with self-similarity. He shows that file system characteristics and user behavior contribute to the prop-

erty. He demonstrates that origin of self-similarity in WWW traffic can be traced to heavy-tailed file transmission times due to the distribution of available file sizes in the web transfers and the heavy tailed silent times due to the influence of user think time. He also suggests that changes in protocol processing and document display are not likely to remove self-similarity of WWW traffic. The work in [19] show that in a realistic client/server environment the degree to which file sizes are heavy tailed directly determines the degree of self-similarity. Also they show that the reliable transmission and flow control mechanisms of TCP(Reno, Tahoe or Vegas) serve to maintain the *long range dependence* structure induced by heavy tailed distributions, in contrast to the unreliable UDP traffic which showed little self-similar characteristics. Active research is going on to determine the causes of *long range dependence* and provide physical explanations and interpretations for such behavior.

Mathematical representations of self-similarity have been provided by many authors. Second order statistics are the statistical properties that capture the burstiness or variability and hence self-similarity is associated with them.

2.2.2 Mathematical definition

Let the time series $X = (X_k : k = 1, 2...)$ represent the number of arrivals in successive non-overlapping intervals of unit time(100ms , 1 sec etc). X is a wide sense stationary stochastic process . The aggregated process $X^{(m)}$ is defined [30]

as

$$X^{(m)}(i) = \frac{1}{m}(X_{(i-1)m+1} + ... + X_{im}) \tag{2.1}$$

where m is the level of aggregation. Then X is called asymptotically self-similar if

$$\lim_{m->\infty} Var(m^{1-H}X^{(m)}) = Var(X) = \sigma^2, 0 < \sigma < \infty, \tag{2.2}$$

and

$$\lim_{m->\infty} r^m(k) = \frac{1}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H}), \tag{2.3}$$

where $r^m = (r^m(k), k \geq 0)$ is the autocorrelation function of the aggregate process and $0 < H < 1$. $X^{(m)}$ is said to be exactly self-similar if $Var(m^{1-H}X^{(m)}) = Var(X) = \sigma^2$ and $r^m = r^m(k) = \frac{1}{2}((k+1)^{2H} - 2k^{2H} + (k-1)^{2H})$ for all m and $0 < H < 1$.

A second-order stationary stochastic process $X = (X_k : k = 0, 1, 2...)$ with autocorrelation function $r(k)$ is said to be *long range dependent* if for some $0 < \beta < 1$,

$$r(k) \sim c_1 k^{-\beta}, k \to \infty \tag{2.4}$$

where c_1 is a positive finite constant and $\beta = 2 - 2H$, H being the Hurst parameter. The autocorrelations decay hyperbolically rather than exponentially fast, implying a non-summable autocorrelation function. The existence of nontrivial correlation at a distant lag is referred to as *long range dependence*. A second-order stationary

9

stochastic process is called *short range dependent* if for some $0 < \rho < 1$,

$$r(k) \sim c_2 \rho^k, k \to \infty \qquad (2.5)$$

where c_2 is a finite positive constant. In contrast to LRD, SRD is characterized by a geometrically decaying and summable autocorrelation function. Hence a self-similar process is long range dependent if its autocorrelation function is non-summable and short range dependent if summable. Equivalent definitions of self-similarity can be found in [3] and [32].

2.2.3   Hurst parameter

Detailed discussions of exactly self-similar and asymptotically self-similar functions can be found in [30]. The degree of self-similarity in a process can be found by estimating the Hurst parameter and it typically depends upon the utilization level of the Ethernet [30]. Higher the value of 'H' more bursty the traffic is. For Poisson traffic H=0.5. Different methods are used for estimating the Hurst parameter namely R/S analysis, aggregated variance method, difference of variance, absolute value of the aggregated series, Higuchi's method, residuals of regression, periodogram method, modified periodogram method, Whittle estimator, etc. are described in [28]. Other measures of burstiness and variability include the 'index of dispersion', 'peak-to-mean ratio', and 'coefficient of variation'. However, 'peak-to-mean ratio' and 'coefficient of variation' measures have been proved to

be unsatisfactory for self-similar traffic [30] .

However, [24] suggests that the Hurst parameter is not a consistent and monotonic indicator of burstiness, and hence LRD property of a traffic source cannot be completely described by Hurst parameter alone. Research evidence also suggests that larger 'H' is associated with smaller queue sizes hence questioning the effectiveness of the 'H' parameter in determining the intensity of LRD.

2.3 Traffic Modeling

2.3.1 Introduction

The Self-similar property of traffic has serious effects on the design, control and performance analysis of high speed data networks. The traditional modeling approaches like the simpler Markovian models fail in the case of self-similar traffic as they often under-estimate the loss probabilities and buffer occupancy levels. This calls for models that would characterize the network traffic and accurately predict the queueing behavior. This section gives an overview of traffic models that were used in the literature focussing on the advantages of Markov models.

2.3.2 Modeling Approaches

Teletraffic modeling aims at evaluating the traffic performance taking into account the network capacity, traffic offered and the performance goals of the system. The input traffic to a network system should be characterized accurately. That is, all the essential statistical characteristics of the traffic seen in measurements should be captured. The traffic characterization should also provide critical insights into the origin of the observed properties, and the models should be computationally feasible (parsimonious modeling). Teletraffic modeling evolved at different stages striving for accurate performance prediction. Willinger and Kihong [20] classified the modeling approaches as *measurement-based traffic modeling*, *physical modeling*, and *queueing analysis*.

2.3.3 Measurement-based traffic modeling

In this approach the data is collected from the physical network and analyzed to detect and identify characteristics. By analyzing the WAN traffic traces Paxson [21] showed that Poisson processes are valid only for modeling the arrival of user sessions such as TELNET connections but fail as accurate models for other WAN arrival processes; he found that WAN packet arrival processes appear better modeled using self-similar processes. Garret and Willinger [8] gave a detailed analysis of a 2 hour long empirical sample of VBR video, and showed the presence of *long range dependence* in the trace. They proposed a source model

which is non-Markovian and stressed that *long range dependence* and heavy tailed marginals need to be considered while modeling VBR traffic. Leland et al. [30] demonstrated that Ethernet LAN traffic is statistically self-similar by analyzing the Bellcore traffic. They suggested the use of self-similar stochastic models like fractional Gaussian noise (fGn), fractional Brownian motion (fBm), fractional autoregressive integrated moving average (FARIMA) models and chaotic maps to fit the Ethernet traffic. Crovella et al. [4] attributed the self-similar property to the WWW transfers in the network and suggested the use of heavy tailed distributions to model such traffic, heavy tailed behavior being inherent to the network traffic. Willinger et al. [31] state that the superposition of many ON/OFF sources (packet trains) whose ON-periods and OFF-periods exhibit the *Noah Effect* (i.e., have high variability or infinite variance) produces aggregate network traffic that features the *Joseph Effect* (i.e., is self-similar or LRD). Here they deal with the traditional ON/OFF source models.

These approaches attempt at matching the traffic data with an assumed statistical model. But consistency with an assumed model does not rule out the possibility of other accurate models that may reflect the properties of the traffic in a better way. Having stated this, these authors succeeded in introducing a new class of models namely the self-similar models and the long-range dependent models. These models fitted to the second-order properties of the traffic very well. However these models failed to give physical explanations on how self-similarity is generated in traffic, nor do they predict behavior over the whole of the design

parameter space.

2.3.4 Physical modeling

This approach aims at models that would explain how traffic is generated in the actual network, that are capable of explaining phenomena such as self-similarity in more elementary terms, and that provide new insights into the dynamic nature of the traffic. For example in MPEG video streams the time duration between successive scene changes exhibits variability at multiple time scales. Research evidence suggest that the video traffic streams exhibit *long range dependence.* Willinger describes this as single source causality. He also addresses structural causality due to the heavy-tailed distribution of file or object sizes. The heavy tailed file size distributions give a physical explanation of network traffic self-similarity. In [31] Willinger et al. established that the *long range dependence* of the aggregated ON/OFF process is determined by the heavy tailedness of the ON or OFF periods thus providing some physical causality of self-similarity. However finer time scale behavior of traffic also has brought attention among the researchers because of variability observed in the packet interarrival times within sessions, flows, or connections.

2.3.5   Queueing analysis

Though the works in the above two categories gave physical interpretations to the observed behavior and accurately captured the second order behavior, they did not attempt at predicting the queueing behavior. The queueing behavior for long range dependent traffic is very different from that for the Poisson or renewal input. Works in this category provide mathematical models of long range dependent traffic which support the analysis and estimation of queueing behavior. These models play a vital role in determining the dimensions of the network. They place performance boundaries on the models by investigating the queueing behavior. The aim of these works is to provide an appropriate model for performance studies. These works can be broadly classified into non-Markov and Markov.

Non-Markov models

The effect of correlations of the input traffic on the queueing systems was studied by Li and Hwang [22]. They observed that queueing performance was dominated by input power in the low frequency band, which is equivalent to long range dependency. But the number of parameters required in such approaches was very high and hence parsimonious models were preferred over highly parameterized models when faced with the task of assigning model parameters in practice. The models like fluid models, transform expand sample (TES), chaotic maps, fBm, fGn and FARIMA were favored by many authors because of their ability to match the

second order properties very well. Eramilli et al. [5] proposed a three param-
eter fractional Brownian motion model (originally adopted by Norros) which is
an exactly self-similar model. They supported the principle of parsimony (also
known as *Occam's Razor*) and gave conditions under which parsimonious models
capture LRD. The conditions they specified include (i) the time scales of interest
in the queueing process coinciding with the scaling region, (ii) the traffic being
aggregated from a large number of independent users, (iii) negligibility of the ef-
fects of flow control. They used the results derived by Norros for determining the
queueing behavior with a deterministic service time. But their results deviated
significantly from reality. The fBm models and other gaussian models had rigid
and restrictive correlation structure. They were not able to effectively capture
the short term correlations whose importance has been demonstrated for queue-
ing in finite length buffers [23]. Ribero et al. [23] formulated multiscale queueing
analysis of long range dependent traffic. They used wavelet-domain independent
Gaussian (WIG) model and multifractal wavelet model (MWM) in their analysis
which had more flexible correlation structures. Grossglauser and Bolot [9] pro-
poses a modulated fluid traffic model and developed a procedure to evaluate the
performance of a finite buffer queue fed with that input.

The results of the above works were difficult to get and harder to apply.
Moreover they failed to predict the queueing behavior accurately and are handi-
capped in not providing physical insight into why they are meaningful on physical
grounds. In the late 90s research works evinced the importance of two other pa-

rameters in performance evaluation, which paved way for Markov models to make an impact in the long range dependent traffic modeling. They are the *critical time scales* and *marginal distributions*.

Critical time scales

Network traffic is known to exhibit multiple time scale behavior. But what are the relevant time scales or time scales of interest for accurate performance prediction? Recent discoveries in [9] suggest that the amount of correlation that needs to be taken into account for performance evaluation depends not only on the correlation structure of the source traffic but also on the time scales specific to the system under study. They found that the time scale associated to a queueing system is a function of the maximum buffer size and that the correlations have negligible impact on performance of the system after what they call correlation horizon (CH). Other authors [6] define the dominant time scale (DTS) as the most probable time scale over which buffer overflow occurs. They also state that the LRD property by itself does not change the buffer distribution, but instead exerts its influence on the value of the DTS by which the buffer behavior is determined. Some papers have also discussed about the importance of CTS in aggregate traffic in backbone networks. Based on the results of large deviation theory Ryu and Elwalid [26] define the CTS of a VBR video source as the number of frame correlations that contribute to the cell loss rate, given the buffer size, link capacity, and the marginal distribution of frame size. They show that second-order behavior at the time scale

beyond the CTS does not significantly affect the network performance. These results imply that a model that can capture correlations up to a finite lag of time would be a good approximation for performance evaluation in finite buffer systems.

Marginal distribution

The work in [9] suggests that the marginal distribution of the traffic arrival process should also be taken into account for accurate loss prediction. So a model used to describe the traffic should match the autocorrelations up to a time lag, keeping the marginals intact. The limitations of using only mean and autocorrelation structure for describing the arrival process in predicting the queuelength distributions has been discussed in [10].

Markov models

The desire for analytical simplicity in models and the quest for physical insights on the traffic generation mechanism triggered the original enthusiasm among researchers in favor of the traditional Markovian modeling approaches. The advantage of Markov models is that it is possible to reuse the well known solution methods developed in the past in order to evaluate the performance of the network. Markov models are known to capture the long range dependencies over a finite range of time scales, which have been proved to be sufficient for performance

evaluation in finite buffer systems. Robert et al. [25] introduced the concept of pseudo long range dependent process which can model the aggregate traffic over several time scales and illustrated that a simple Markov modulated model based on the theory of near complete decomposability (developed by Courtois [2]) can approximate LRD traffic. Feldmann and Whit [7] developed a fitting algorithm for approximating a heavy tailed or long tailed distribution using a finite mixture of exponentials called hyper-exponentials. Though the hyper-exponentials have an exponential tail they can approximate heavy-tailedness in the regions of primary interest. This made queueing solution methods easier and motivated other researchers to develop fitting algorithms on similar lines.

Anderson et al. [1] showed that the superposition of several two-state Markov modulated Poisson process can model the self-similar behavior over several time scales of interest. Their queueing results based on the matrix analytical approach developed by Neuts [18] did not accurately predict the queueing behavior suggesting the importance of fitting the marginals. They achieve a very good fit on the covariances over several time scales by using a fitting algorithm based on weighted sum of exponentials.

A similar fitting algorithm was used in [32] to fit the variance of their MMPP model over several time scales. Again their fitting of second order statistics alone was not sufficient to predict the queueing behavior accurately. However Kasahara [13] concluded that variance fitting method seems to be enough to predict the loss behavior of finite queueing systems with LRD input when appropriate

time scale is considered. Valados et al. [27] support the concept of critical time scales and proposed an MMPP model that would match the covariance structure of the process as well as the marginal distribution. They obtained very good results in estimating the loss probabilities at higher utilization levels but under light loads their results deviated. Though the aforementioned works provided analytically simpler solutions and achieved a certain degree of success in predicting the queueing behavior, a solution form providing deep insight into the nature of the system was still missing.

Systems with long range dependencies can be divided into subsystems such that the interaction between certain subsystems is smaller than the interactions within the subsystems [12]. This property is called NCD and was shown by [12] to be useful to derive analytical solutions, based on linear algebraic queueing theory (LAQT) which gives flexibility in generating arrival processes with arbitrary marginal and correlation structures. This method yielded the much needed critical insight.

2.4 Linear Algebraic queueing theory

2.4.1 Matrix exponential distribution

A matrix exponential distribution is defined as a probability distribution with representation $(\mathbf{p}, \mathbf{B}, \boldsymbol{\epsilon})$ [14]. i.e.,

$$F(t) = 1 - \mathbf{p} \, \exp(-\mathbf{B}t) \, \boldsymbol{\epsilon}', \quad t \geq 0, \tag{2.6}$$

where $\mathbf{p}$ is the starting vector for the process, $\mathbf{B}$ is the process rate operator or the progress rate matrix which must be non-singular, and $\boldsymbol{\epsilon}'$ (the transpose of $\boldsymbol{\epsilon}$) is the summing operator. The order of the representation is indicated by the dimension of the $\mathbf{B}$ matrix, and the degree of the distribution $F(t)$ is the minimal order of all its representations. Its probability density function is defined as

$$f(t) = \frac{dF(t)}{dt} = \mathbf{p} \, \exp(-\mathbf{B}t) \, \mathbf{B} \, \boldsymbol{\epsilon}' \tag{2.7}$$

The nth moments satisfy the following

$$E[X^n] = \int_0^\infty x^n f(t) \, dt = n! \mathbf{p} \mathbf{V}^n \boldsymbol{\epsilon}' \tag{2.8}$$

21

where $\mathbf{V} = \mathbf{B}^{-1}$. The Laplace-Stieltjes Transform of $f(t)$ is given by

$$\mathbf{B}^*(s) = \int_0^\infty \exp^{-st} f(t)\, dt = \mathbf{p}(\mathbf{I} + s\mathbf{V})^{-1}\boldsymbol{\epsilon}' = \mathbf{p}(\mathbf{B} + s\mathbf{I})^{-1}\mathbf{B}\boldsymbol{\epsilon}' \qquad (2.9)$$

Matrix exponential distributions have rational Laplace-Stieltjes transform and are more general than the Phase-type (PH) distributions defined by Neuts [18]. ME distributions place fewer constraints on its representation. Although the class of second degree matrix exponential distributions are equivalent to the physically based phase-type distributions, higher degree representations may not have physical representation. Phase type distributions are, in fact, a strict subset of matrix exponential distributions.

The class of distributions with rational Laplace-Stieltjes transforms is dense in the set of all distributions, which means that any density function can be approximated arbitrarily closely by a density with a rational transform. Some of such distributions are exponential, Erlangian, Coxian, hypoexponential, hyperexponential, Marie, and mixtures or convolutions of these distributions. Distributions belonging to the above class have probabilistically interpretable components and have close relationship with Markov chains. The advantage of using matrix exponential distributions is that higher order moments can be matched using Appie van de Liefvoort's algorithm [29] and they can be represented in different canonical forms through the use of similarity transforms. The LAQT solution method

does not depend on the process representation and hence any representation can be chosen while modeling.

2.4.2 Moment matching

Appie van de Liefvoort's algorithm [29] can be used to match the moments of the distribution. From the set of power moments $E[X^n]$ of a continuous distribution $F(t)$ an ME distribution $(\mathbf{p}, \mathbf{B}, \boldsymbol{\epsilon})$ can be generated. Let,

$$r_n = \frac{E[X^n]}{n!} \tag{2.10}$$

be the set of normalized or reduced moments of the distribution. Applying the algorithm we can generate,

$$\mathbf{p} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} r_1 & r_1 \\ (r_2 - r_1^2)/r_1 & (r_3 - 2r_1 r_2 + r_1^3)/(r_2 - r_1^2) \end{bmatrix}, \quad \boldsymbol{\epsilon}' = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{2.11}$$

using the first three moments. This representation has the prescribed moments and a rational Laplace-Stieltjes transform. The boundary conditions of the power moments have been discussed in [16]. For a third order representation, the first five moments are mapped into $(\mathbf{p}, \mathbf{B}, \boldsymbol{\epsilon})$

$$\mathbf{p} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{\epsilon}' = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \tag{2.12}$$

$$\mathbf{V} = \begin{bmatrix} r_1 & r_1 & 0 \\ (r_2 - r_1^2)/r_1 & (r_3 - 2r_1r_2 + r_1^3)/(r_2 - r_1^2) & r1 \\ 0 & -(r_2^3 - 2r_1r_2r_3 + r_3^2 + r_4r_1^2 - r_4r_2)/(r_1^2 - r_2)^2r_1 & \beta/\gamma \end{bmatrix} \tag{2.13}$$

where

$$\beta = -r_2^4 r_1 + 3r_2^2 r_1^2 r_3 - 2r_1r_2r_3^2 - 2r_2r_1^3 r_4 + 2r_2^2 r_1 r_4 - r_3^2 r_1^3 + r_3^3 + 2r_3r_4r_1^2$$

$$- 2r_3r_2r_4 + r_5r_1^4 - 2r_5r_1^2 r_2 + r_5r_2^2,$$

$$\gamma = (r_2^3 - 2r_1r_2r_3 + r_3^2 + r_4r_1^2 - r_4r_2)(r_1^2 - r_2)$$

2.4.3 Sequence of ME intervals

If $T_1, T_2, T_3, \ldots$ is a sequence of ME random variables then the joint probability density function over any finite sequence inter-event times is given by

$$f_{T_1,T_2,...,T_n}(t_1,...,t_n) = \boldsymbol{\pi}(0)\,exp(-\mathbf{B}t_1)\,\mathbf{L}...exp(-\mathbf{B}t_n)\,\mathbf{L}\boldsymbol{\epsilon}' \qquad (2.14)$$

where $\boldsymbol{\pi}(t)$ is a vector representing the internal state of the process at time t and $\mathbf{L}$ is the event rate matrix. If the process is renewal then $\mathbf{L} = \mathbf{B}\boldsymbol{\epsilon}'\mathbf{p}$ where $\mathbf{p}$ is the starting vector for the process, the rank of $\mathbf{L}$ being 1.

If L is of rank greater than 1, then the process will exhibit auto dependence. The covariance of a sequence of MEs [17, 18] assuming stationarity is given by,

$$cov[X_n, X_{n+k}] = \mathbf{p}\mathbf{V}(\mathbf{Y})^k\mathbf{V}\boldsymbol{\epsilon}' - (\mathbf{p}\mathbf{V}\boldsymbol{\epsilon}')^2 \qquad (2.15)$$

and the variance,

$$var[X_0] = 2\mathbf{p}\mathbf{V}^2\boldsymbol{\epsilon}' - (\mathbf{p}\mathbf{V}\boldsymbol{\epsilon}')^2 \qquad (2.16)$$

where $\mathbf{V} = \mathbf{B}^{-1}$ and $\mathbf{Y} = \mathbf{V}\mathbf{L}$ such that $\mathbf{Y}\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}'$ and $\mathbf{p}\mathbf{Y} = \mathbf{p}$. Thus $\boldsymbol{\epsilon}'$ is the right eigenvector of $\mathbf{Y}$ with eigenvalue 1 and $\mathbf{p}$ is the left eigenvector of $\mathbf{Y}$ with eigenvalue 1. The value of $\mathbf{p}$ is assumed to be unique and its existence is guaranteed if 1 is the largest eigenvector of $\mathbf{Y}$. The autocorrelation is obtained by dividing the covariance by the variance,

$$Autocorr[X_n, X_{n+k}] = \frac{\mathbf{pVY}^k\mathbf{V}\boldsymbol{\epsilon}' - (\mathbf{pV}\boldsymbol{\epsilon}')^2}{2\mathbf{pV}^2\boldsymbol{\epsilon}' - (\mathbf{pV}\boldsymbol{\epsilon}')^2} \tag{2.17}$$

If $\mathbf{L} = \mathbf{B}\boldsymbol{\epsilon}'\mathbf{p}$ then $\mathbf{Y}$ is of rank 1 and the process becomes renewal and hence $cov[X_n, X_{n+k}] = 0$.

2.4.4   Introducing correlations

Auto-correlations can be arbitrarily introduced into a renewal process preserving the marginal distribution the process by using Mitchell's method. The progress rate matrix of the arrival process ($\mathbf{B}$) describes what happens before the arrival event and the event rate matrix describes ($\mathbf{L}$) what happens at the time of an arrival. $\mathbf{L}$ can be altered from representing a renewal process to represent a semi-Markov arrival process, which is correlated. The starting vector $\mathbf{p}$ and the progress rate matrix $\mathbf{B}$ are kept unchanged from the renewal process representation. An $\mathbf{L}$ can be chosen such that $\mathbf{p}$ and $\mathbf{B}$ remain unchanged,

$$\mathbf{L} = \beta(\mathbf{B}\boldsymbol{\epsilon}'\mathbf{p} - \mathbf{B}) + \mathbf{B} \tag{2.18}$$

$$\mathbf{L} = (1 - \gamma)(\mathbf{B}\boldsymbol{\epsilon}'\mathbf{p} - \mathbf{B}) + \mathbf{B} \tag{2.19}$$

where $\beta = 1 - \gamma$ is called the persistence of the process, and as $\beta \to 0$, the autocorrelation exhibited by the point process increases. This method can be used to match the decay of the desired autocorrelation structure rather than matching the entire autocorrelation structure of the process, which Mitchell has found out to be more effective.

2.4.5 Kronecker products and Hat spaces

Kronecker products [14] are used to combine processes operating in different spaces. It is a way to preserve independence among process in different spaces, in which disjoint operator spaces are embedded into the direct product space. For example in a ME/ME/1/N queueing system the arrival process $< \mathbf{B}_a, \mathbf{L}_a >$ and the service process $< \mathbf{B}_s, \mathbf{L}_s >$ operate in two disjoint spaces namely the arrival space and the service space. Their independence is preserved by using the Kronecker product. The Kronecker product is represented by the symbol $\otimes$. The Kronecker product of two matrices $\mathbf{K}_1$ (operating on objects in space 1) and $\mathbf{K}_2$ (operating on objects in space 2) is given by,

$$\mathbf{K} = \mathbf{K}_1 \otimes \mathbf{K}_2 = \begin{bmatrix} (\mathbf{K}_1)_{11}\mathbf{K}_2 & (\mathbf{K}_1)_{12}\mathbf{K}_2 & (\mathbf{K}_1)_{13}\mathbf{K}_2 \\ (\mathbf{K}_1)_{21}\mathbf{K}_2 & (\mathbf{K}_1)_{22}\mathbf{K}_2 & (\mathbf{K}_1)_{23}\mathbf{K}_2 \end{bmatrix} \tag{2.20}$$

where $\mathbf{K}_1$ is of size 2×3 and $\mathbf{K}_2$ is of size 2×2. The Kronecker product is

not commutative or symmetric. That is, $\mathbf{K}_1 \otimes \mathbf{K}_2 \neq \mathbf{K}_2 \otimes \mathbf{K}_1$ although the two representations are equivalent provided order is consistent within a formula. To hide the ordering and simplify visually, hat spaces are introduced. For example if there are N spaces, then

$$\widehat{\mathbf{K}}_1 = \mathbf{K}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{I}_N$$

$$\widehat{\mathbf{K}}_2 = \mathbf{I}_1 \otimes \mathbf{K}_2 \otimes I_3 ... \otimes \mathbf{I}_N \qquad (2.21)$$

$$\widehat{\mathbf{K}}_N = \mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{K}_N$$

where $\mathbf{I}_i$ is the identity matrix of dimensions $m_i \times m_i$ and $\mathbf{K}_i$ is the matrix of that dimension. The property,

$$\widehat{\mathbf{K}}_1 \cdot \widehat{\mathbf{K}}_2 = \mathbf{K}_1 \otimes \mathbf{K}_2 = \widehat{\mathbf{K}}_2 \cdot \widehat{\mathbf{K}}_1 \qquad (2.22)$$

satisfies the independence of spaces condition. For the matrix exponentials operating in N independent spaces, the resultant matrices and vectors are given by,

Matrix $\mathbf{B}_i$:

$$\widehat{\mathbf{B}}_1 := \mathbf{B}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{I}_N$$

$$\widehat{\mathbf{B}}_2 := \mathbf{I}_1 \otimes \mathbf{B}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{I}_N \qquad (2.23)$$

$$\widehat{\mathbf{B}}_N := \mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{B}_N$$

Matrix $\mathbf{L}_i$:

$$\widehat{\mathbf{L}}_1 := \mathbf{L}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{I}_N$$

$$\widehat{\mathbf{L}}_2 := \mathbf{I}_1 \otimes \mathbf{L}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{I}_N \qquad (2.24)$$

$$\widehat{\mathbf{L}}_N := \mathbf{I}_1 \otimes \mathbf{I}_2 \otimes \mathbf{I}_3 ... \otimes \mathbf{L}_N$$

Starting vector $\mathbf{p}$:

$$\widehat{\mathbf{p}}_1 \cdot \widehat{\mathbf{p}}_2 \neq \mathbf{p} \qquad (2.25)$$

$$\mathbf{p} = \mathbf{p}_1 \otimes \mathbf{p}_2 = \mathbf{p}_1 \widehat{\mathbf{p}}_2 = \mathbf{p}_2 \widehat{\mathbf{p}}_1 \qquad (2.26)$$

Summing vector $\boldsymbol{\epsilon}'$:

$$\widehat{\boldsymbol{\epsilon}'}_1 \cdot \widehat{\boldsymbol{\epsilon}'}_2 \neq \boldsymbol{\epsilon}' \qquad (2.27)$$

$$\boldsymbol{\epsilon}' = \boldsymbol{\epsilon}'_1 \otimes \boldsymbol{\epsilon}'_2 = \widehat{\boldsymbol{\epsilon}'}_2 \boldsymbol{\epsilon}'_1 = \widehat{\boldsymbol{\epsilon}'}_1 \boldsymbol{\epsilon}'_2 \qquad (2.28)$$

Chapter 3

Mode Extraction

3.1 Introduction

The packet losses occurring in a queueing system due to bursty traffic can be explained by the concept of weak stability. The traffic arrival mechanism can be modal. Modality in traffic can explain the weakly stable conditions observed in the queues and also can characterize the observed burstiness. This chapter explains the concept of modes in bursty traffic and the different methods to extract them from traffic trace data. The major features of the extraction process are also discussed. This chapter also deals with the properties of different modes and their effect on a queueing system providing necessary insight into the actual nature of the traffic arrival mechanism. The extraction process is illustrated through analysis of two sets of traffic traces described in section 3.3.

3.2   Modes and Weak stability

A premise of this work is that Ethernet traffic can be adequately represented by a continuous time Markov chain (cpMc) model that captures the dependencies and burstiness observed in the traffic, and that this traffic model can be used to predict the queueing behavior of networking systems that handle this traffic.

Burstiness in traffic can be characterized by a modal cpMc. A modal cpMc can be defined as a cpMc whose state transitions are nearly completely decomposable, with decomposition classes each defining a particular arrival rate, interarrival time distribution and/or dependency. These classes are called modes because to a general observer the traffic pattern appears different whenever the nearly-decomposable process jumps from class to class. To appear bursty one or more modes must have a mean arrival rate which is significantly greater than the mean arrival rate taken over all modes.

The concept of weak-stability, discussed by Jelenkovich [11], suggests that with respect to traffic, a stable queueing system can be either weakly stable or strictly stable. When it is weakly stable the arrival rate of one or more modes exceed the service capacity, though the overall service rate is greater than the overall arrival rate. As the modal traffic moves from mode to mode, the system moves from stable to unstable even though the overall system is stable. On the other hand, when the service capacity is greater than the arrival rate of the fastest mode, the system is strictly stable. Figure 3.1 explains the concept of weak

stability. The modes with mean arrival rate higher than the mean service rate are unstable and the rest of the modes in the traffic are stable. As the service rate is adjusted upward, unstable modes can become stable.
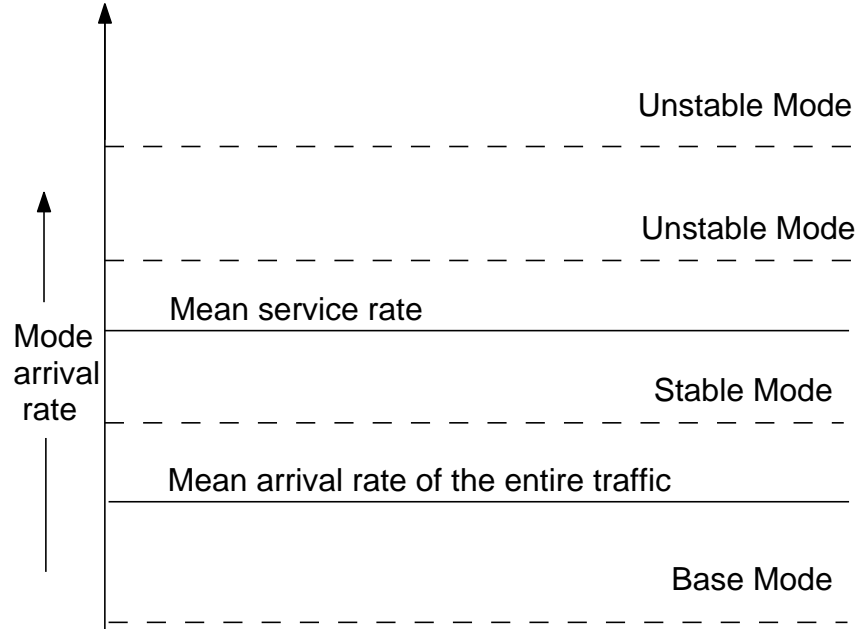

Figure 3.1: Modes in traffic

3.3 Tactics

One of the goals of this thesis is to identify and extract the instabilities in the form of unstable modes from the traffic. Efficient methods will be developed to extract these modes from publicly available traffic traces. The extraction of unstable modes from the traffic leaves behind the base mode. It is possible that the modal structure extends even into the base mode (there might be slower sub-modes within the base mode). But the extraction process will not attempt at extracting these sub-modes as it is out of scope of this thesis.

Based on the properties of the extracted modes an analytical model will be built. Two stage hyperexponential distributions with physically interpretable phases will be used to model the interarrival time distributions of the modes. This makes the overall model simpler and easier to analyze, providing insight. The solution techniques also become simple with the use of hyperexponential distributions. The mode durations will be modeled using exponential distributions. Mitchell's correlation model can be used to introduce finite lag autocorrelations in the interarrival times of modes where autocorrelations are appropriate. The correlation model is a single parameter model and cannot match the exact correlation structure but can provide long but finite lag correlations.

The main strength of the multi-modal analytical model is that the instabilities will be modeled very well. So the model is expected to perform well at lower utilizations at which the unstable modes play a vital role in determining the packet losses. The weakness comes in the form of the poor modeling of marginal distribution of the base mode. The sub-modes are not modeled and this might lead to poor performance at high utilizations at which the role of base mode is vital in determining the queue occupancy. The base mode might well have slowly decaying correlation structure which cannot be matched by Mitchell's model. This can be a deficiency in the model. It is possible that if the sub-modes are modeled properly, the long scale interactions between them will automatically model the correlations. It should be noted that memory is retained whenever there is a transition from faster time scale to slower time scale and vice versa.

The queueing system will be modeled using exponentially distributed service times and finite buffer. Tractable solution techniques will be developed and verified. The performance of the multi-modal model will be estimated by comparing the packet loss probabilities obtained from the multi-modal model with those obtained from the original trace driven simulations. To demonstrate the strength of the multi-modal model, a comparison will be done with a published model.

3.4 Traffic traces used in extraction

Two traffic traces, namely the famous October and August Bellcore traces, were used to test for the presence of modes. The following section describes the properties of these traces.

3.4.1 Bellcore October and August traces

The following description of the traffic traces is from

http://ita.ee.lbl.gov/html/contrib/BC.html:

The trace BC-pAug89 began at 11:25 on August 29, 1989, and ran for about 3142.82 seconds (until 1,000,000 packets had been captured). The trace BC-pOct89 began at 11:00 on October 5, 1989, and ran for about 1759.62 seconds(until 1,000,000 packets had been captured).

The trace captured all Ethernet packets. The files BC-pOct89.TL and BC-pAug89.TL are ASCII-format tracing data, consisting of one 20-byte line per Ethernet packet arrival. Each line contains a floating-point time stamp (representing the time in seconds since the start of a trace)and an integer length (representing the Ethernet data length in bytes). Although the times are expressed to 6 places after the decimal point, giving the appearance of microsecond resolution, the hardware clock had an actual resolution of 4 microseconds. The length field does not include the Ethernet preamble, header, or CRC; however, the Ethernet protocol forces all packets to have at least the minimum size of 64 bytes and at most the maximum size of 1518 bytes. 99.5% of the encapsulated packets carried by the Ethernet PDUs were IP. All traces were conducted on an Ethernet cable at the Bellcore Morristown Research and Engineering facility, building MRE-2. At that time, the Ethernet cable nicknamed the "purple cable" carried not only a major portion of their Lab's traffic but also all traffic to and from the internet and all of Bellcore. The records include all complete packets (the monitor did not artificially "clip" traffic bursts), but do not include any fragments or collisions. These samples are excerpts from approximately 300 million arrivals recorded; the complete trace records included Ethernet status flags, the Ethernet source and destination, and the first 60 bytes of each encapsulated packet (allowing identification of higher-level protocols, IP source and destination fields,

and so on).

3.5   Extraction Procedure

As described in section 3.2, in stable systems the overall mean arrival rate for the input traffic will be less than the mean service rate. But in weakly stable systems there may be one or more modes, rarely entered, having mean arrival rate greater than that of the service rate of the system. These modes are called as unstable modes, and if there is more than one, they will differ in their mean arrival rates.

The aim of the extraction process is to identify and classify modes of traffic which are characterized by different arrival rates. A queue and server is used to determine these modes by examining the rise and drop of queue lengths. If a exponential server is used, the rate at which the queue length drops and rises is determined by the local variations in service rates. Due to this there is a very good chance that a few arrivals belonging to a particular mode may not get counted towards that mode. So a deterministic server is used to eliminate this problem of local variations in service rates. Service distributions does affect the extraction process but deterministic service will allow us to identify the boundaries of the modes clearly without ambiguity.

An analysis of the queue length evolution in a 'trace/D/1' system with pOct.TL Bellcore trace as the input stream provides deep insight on the nature of

the modes. At higher utilizations the system may exhibit a multiple of unstable modes while at sufficiently low utilizations it may exhibit just one which is the fastest mode. Analysis was performed at different utilizations to be able to identify and possibly extract data about each of the unstable modes individually. Figures 3.2, 3.3 and 3.4 zoom in on the queue length evolution at different utilizations, to illustrate the effect of mode 1, the fastest mode observed.
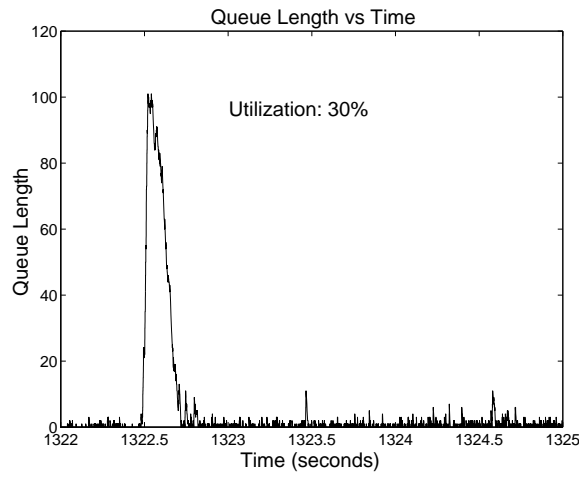

Figure 3.2: Queue length evolution: $\rho = 0.3$
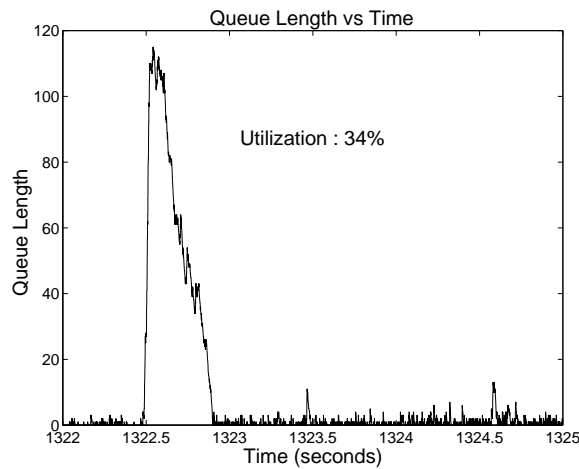
Figure 3.3: Queue length evolution: $\rho = 0.34$

From the figures we can observe the following:
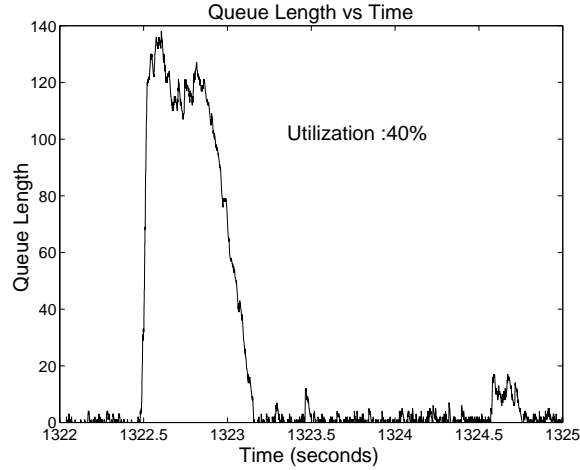
Figure 3.4: Queue length evolution: $\rho = 0.4$

1. The start time of the burst of arrivals due to the mode is not affected by changes in utilization and the start is marked by consecutive arrivals. For example, for the mode depicted in the figures the start of the burst is around 1322.49 seconds.

2. The peak queue length reached by the burst belonging to the mode is considered to be the end of the burst since it is clear that arrivals have slowed. The peak seems to attain higher values with higher utilizations. This is because, more arrivals add to the peak of the burst due to lower service rates at higher utilizations.

3. The region between the start and the peak constitutes the interval of arrivals in that mode. The mean rate of queue length increase in this region is given by $\lambda - \mu$ where λ is the mean arrival rate of the mode and μ is the mean service rate.

4. After the peak the queue length drops, as the mode switches to the base

38

mode. The slope reflects the arrivals in the base mode which have lower arrival rates (less than the mean arrival rate of the entire traffic stream).

5. From figure 3.4 we can observe a relatively slower mode becoming prominent at 1324.6 seconds. Similarly various modes become evident at different instants as the utilization is increased.

6. All the observed modes behave in a similar manner but the rate at which they reach the peak is dependent on their mean arrival rate.

3.5.1 Extraction Methodologies

Packets contributing to the modal behavior can be captured by observing the regions where the queue length exceeds a certain threshold queue size at different utilizations. The packets belonging to the fastest mode can be extracted by fixing the utilization low. But the above observations suggest that, as the utilization is increased the adjoining packets with relatively slower rate contribute to the mode and so the mode durations see an increase. The new packets joining the mode may or may not cause a change (decrease) in arrival rate of the mode, but one may not deny their association to that particular mode.

As suggested by Dr. Wallace, the extraction process can be performed step by step by visual interpretation (by viewing the queue evolution over time at different utilizations). The traffic stream was provided as an input to a sin-

gle server deterministic queueing system with infinite buffer, deterministic service times enabling clear identification of modes. Setting the utilization to a very low value, the time instants of the start and end of the fastest mode were traced, the start being indicated by increase in queue length beyond a threshold, the end being indicated by queue length dropping below the peak. As explained before there may be arrivals belonging to this mode at an adjacent interval not contributing significantly to the buffer increase because of low utilization. There can also be adjacent bursts belonging to the same mode that may merge as utilization is increased and its possibility cannot be ruled out because one cannot expect queue lengths to increase continuously without a few services happening. The utilization was increased until a level after which newer modes began to appear. The characteristics of the bursts belonging to this mode (burst durations, inter-arrival times, mean arrival rate) were obtained from trace driven simulation. In the simulation, whenever there is a burst, all the essential characteristics of the arrivals belonging to the burst were monitored and buffered. From the buffered data the marginal distribution of the interarrival times was estimated. Extracting the fastest mode from the trace leaves behind the relatively slower modes. Then the same experiment was performed to extract the relatively slower modes.

This method had some disadvantages.

1. Observing the queueing behavior over the entire traffic duration is cumbersome.

2. The extraction method entirely depends upon visual interpretation which is prone to errors.

3. The method is extremely cumbersome if the traffic has a large number of modes.

4. The communication between different modes is difficult to establish.

5. There is no standard method to determine the buffer threshold level to be set for extracting different modes.

Because of these disadvantages a better strategy was developed. This method is based on the fact that bursts belonging to a particular mode will cause an increase in the queue length at the same rates $(\lambda - \mu)$. This is applicable to all utilizations. Start of a burst was identified whenever there is a sudden rise in queue length triggered by 5 or more consecutive arrivals without a service happening. The end of a burst was identified by two or more consecutive services before the next arrival (There was at least a single service occurring during the presence of a mode and the bursts would have parted into smaller insignificant lengths had single service been used as end of burst). Since short bursts can also occur as a natural effect of randomness in the base mode, bursts with arrival count less than

10 were not taken into account (Poisson arrivals were determined to have very low chance of bursts of more than 10). Again all the arrivals in the burst were buffered to determine their characteristics.

The main features of this method are:

1. Most of the modes can be captured by fixing the utilization high, say 95%, because the start of the burst belonging to a particular mode will be the same regardless of utilization. Also, the peak queue length reached due to the mode will be the highest at this utilization, so all the arrivals associated with that mode can be captured.

2. There is no discrepancy in setting the buffer threshold levels and no visual interpretation or manual intervention is needed in extraction of modes. The process is automated.

3. The bursts extracted by this method have different mean interarrival times. The bursts can be grouped together based on the mean interarrival times, and bursts belonging to the same group form a mode. The mode having the smallest mean interarrival time is the fastest mode and the one having the highest mean interarrival time is the slowest mode.

4. From the simulation, properties of modes like mean burst duration in a mode, mean arrival rate for a mode, mean burst interarrival time, squared coefficient of variation of interarrival times within a mode, mean burst size (in bytes) in a mode can be determined.

5. After extracting the characteristics of the potentially unstable modes, characteristics of the base mode can be determined by looking at the trace data remaining after removing all the packets constituting the unstable modes.

6. Burst of two different modes do not overlap and hence the properties of a mode are distinct.

7. Communication between different modes in the traffic can be easily determined. By observing the mode arrival instants in the traffic, the mode that existed before the current mode and the mode that will appear after the current mode can be easily determined. Based on this, the rate at which one mode jumps to another mode can be estimated.

3.6 Windowing method

As a third option, the modal nature of the traffic can also be identified by a windowing method. In this method the traffic stream is spilt into windows of equal size and the average arrival rate in those windows is calculated. This method is simple in the sense that a queueing system need not be used to determine the presence of modes. The histogram plot (Figure 3.5) of the arrival rates in every window gives an idea of the unstable regions in the traffic. Adjacent windows having the same arrival rate can be merged and can be counted as a burst belonging to the respective mode. One should be careful in the selection of the window
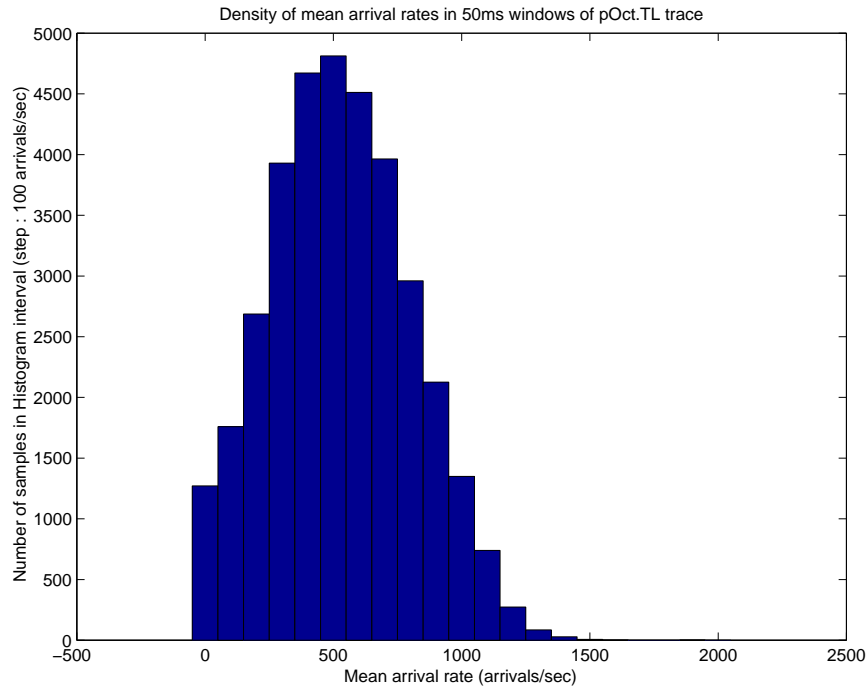
Figure 3.5: Density of mean arrival rate in 50ms windows of pOct.TL trace

size such that the arrivals due to the base mode do not make contribution to the bursts of the faster modes. This can be overcome by using a binary search algorithm (not implemented) within the windows adjacent to each other. The algorithm recursively looks for sections in a window which have the arrival rate of the previous window. If the rates are same, then this section is merged with the previous window. This method has not been used in this work to extract the modes, but provides useful information about the unstable regions in the traffic.

3.7 Properties of modes

The properties that characterize a mode are mean arrival rate, marginal distribution of interarrival times, coefficient of variation and autocorrelation of interarrival times, and the mean duration of bursts belonging to that mode. These properties are used in building an analytical model described in the next chapter. By using the extraction process described above, many modes having a wide range of arrival rates were observed. These modes were grouped in discrete bins based on their mean interarrival times. The grouping yielded a total of nine modes including the stable base mode. Tables 3.1 and 3.2 list the mean interarrival time, coefficient of variation of the interarrival times, and the mean burst duration for different modes found in the pOct.TL and pAug.TL trace data respectively. Mode 1 is the fastest mode and mode 9 represents the base mode. The mean interarrival time for the entire pOct.TL trace is $0.00176s$ and for the pAug.TL trace is $0.003143s$. The following discussions are with respect to the pOct.TL trace. Graphical plots representing the properties of pAug.TL can be found in the appendix.

3.7.1 Interarrival time distributions

This part deals with the interarrival time distribution of different modes. Figure 3.6 shows the sample interarrival time distribution for the entire traffic trace. It can be observed that the marginal distribution has a long tail. Two prominent peaks can be seen around $0.1ms$ and $1ms$ implying that the traffic consists mostly

Modes	Range (s)	Mean interarrival time (s)	C^2	Mean burst duration (s)
1	0.0003-0.0004	0.000317	1.427585	0.026185
2	0.0004-0.0006	0.000560	0.671417	0.012560
3	0.0006-0.0007	0.000664	0.490000	0.017110
4	0.0007-0.0008	0.000752	0.450612	0.021943
5	0.0008-0.0009	0.000842	0.425600	0.023674
6	0.0009-0.0010	0.000937	0.508573	0.024130
7	0.0010-0.0012	0.001060	1.214025	0.022000
8	0.0012-0.0017	0.001362	4.669000	0.033270
9	>0.0017	0.001824	3.249141	0.768197

Table 3.1: Modes of pOct.TL

Modes	Range (s)	Mean interarrival time (s)	C^2	Mean burst duration (s)
1	0.0005-0.0009	0.000832	0.563744	0.024169
2	0.0009-0.0010	0.000957	0.576685	0.035730
3	0.0010-0.0011	0.001052	0.578358	0.039754
4	0.0011-0.0012	0.001152	0.582912	0.048140
5	0.0012-0.0013	0.001246	0.594057	0.044846
6	0.0013-0.0014	0.001346	0.564025	0.043820
7	0.0014-0.0016	0.001485	0.538976	0.047230
8	0.0016-0.0030	0.001765	0.811398	0.045357
9	> 0.0030	0.003296	3.132113	1.384005

Table 3.2: Modes of pAug.TL

of arrivals with shorter interarrival times around these peaks, which is very different from that of Poisson arrivals having exponential interarrival distribution. The base mode interarrival time distribution shown in Figure 3.7 is also heavy tailed with similar peaks observed in the distribution plot of the original trace.

Figure 3.8 and 3.9 show the interarrival time distribution for modes 1 to 8. The plots show a great deal of variation in the number of arrivals in these modes and also the relative prominence of two peaks around $0.1ms$ and $1ms$. In mode 1 only the peak at $0.1ms$ is prominent, signifying a higher mean arrival rate. The number of arrivals in modes 2 and 3 are higher than in the mode 1. Also
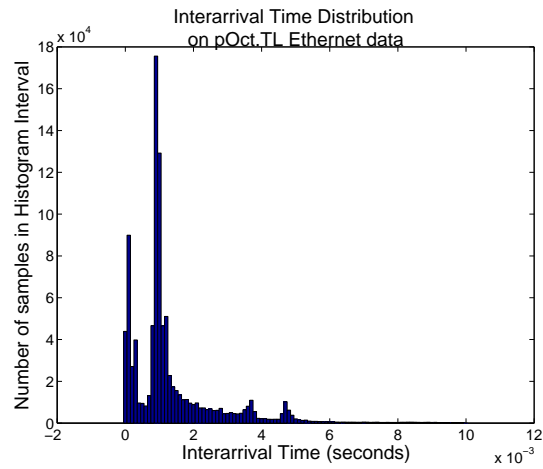
Figure 3.6: Interarrival time distribution: pOct.TL
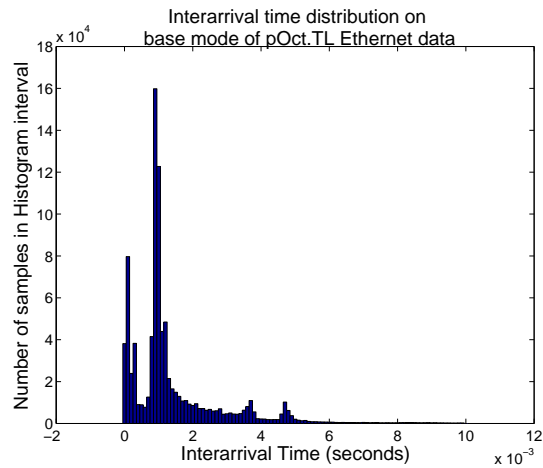
Figure 3.7: Interarrival time distribution: Base mode extracted from pOct.TL

the arrivals with interarrival times around $1ms$ gain prominence in modes 2 and 3 explaining the relative decrease in the mean arrival rate in these modes. This trend is observed also in rest of the modes.
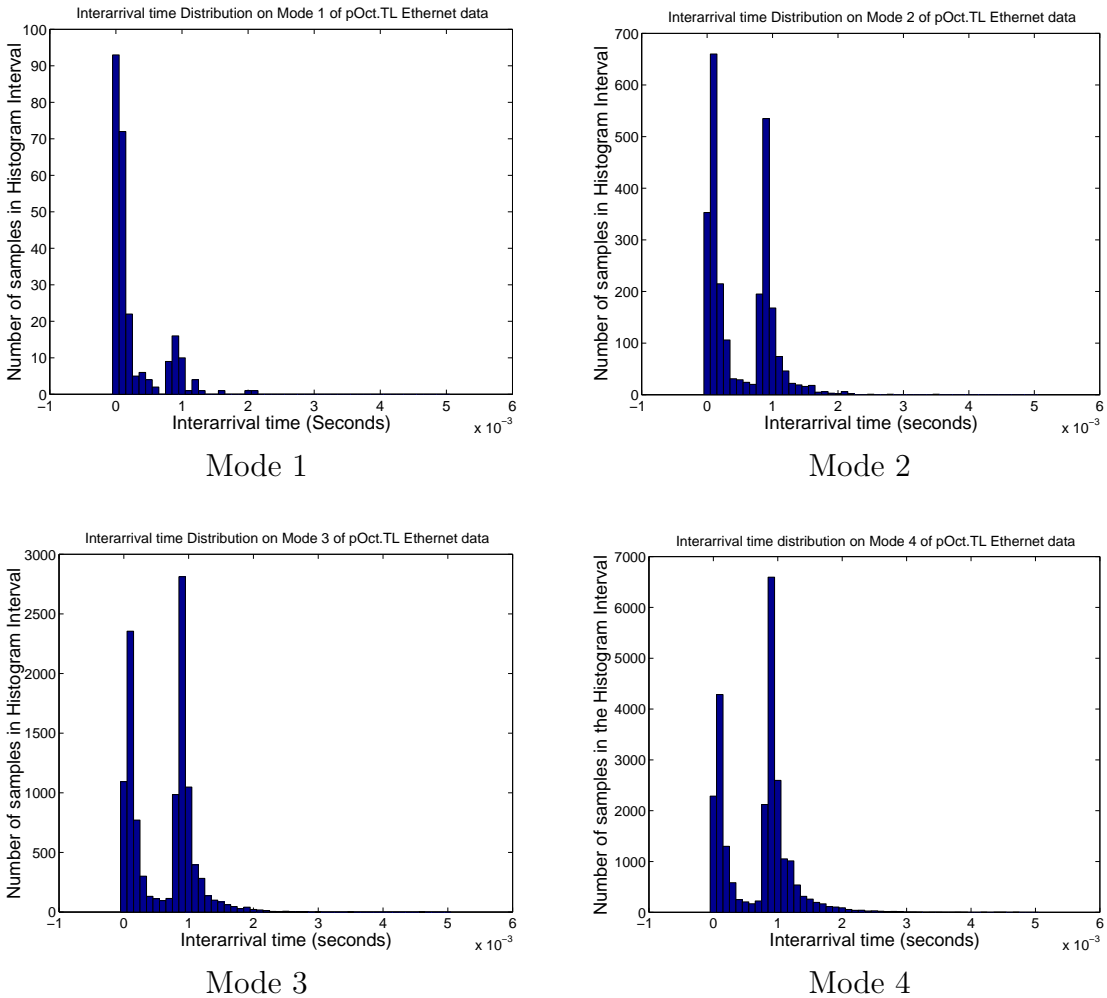
Figure 3.8: Interarrival time distributions-Modes 1-4 of pOct.TL trace

3.7.2 Coefficient of variation and burst duration

Two other features, namely the coefficient of variation (C^2) and the burst duration also distinguish the modes. The coefficient of variation is 1 for exponential distribution. From the tables 3.1 and 3.2 we can see that the value of C^2 is notably different for each of the modes. Modes with values of C^2 greater than 1 can be represented using a hyperexponential (sum of exponentials) distribution and those with C^2 less than 1 can be represented using a hypoexponential (convolu-

Mode 5

Mode 6

Mode 7

Mode 8
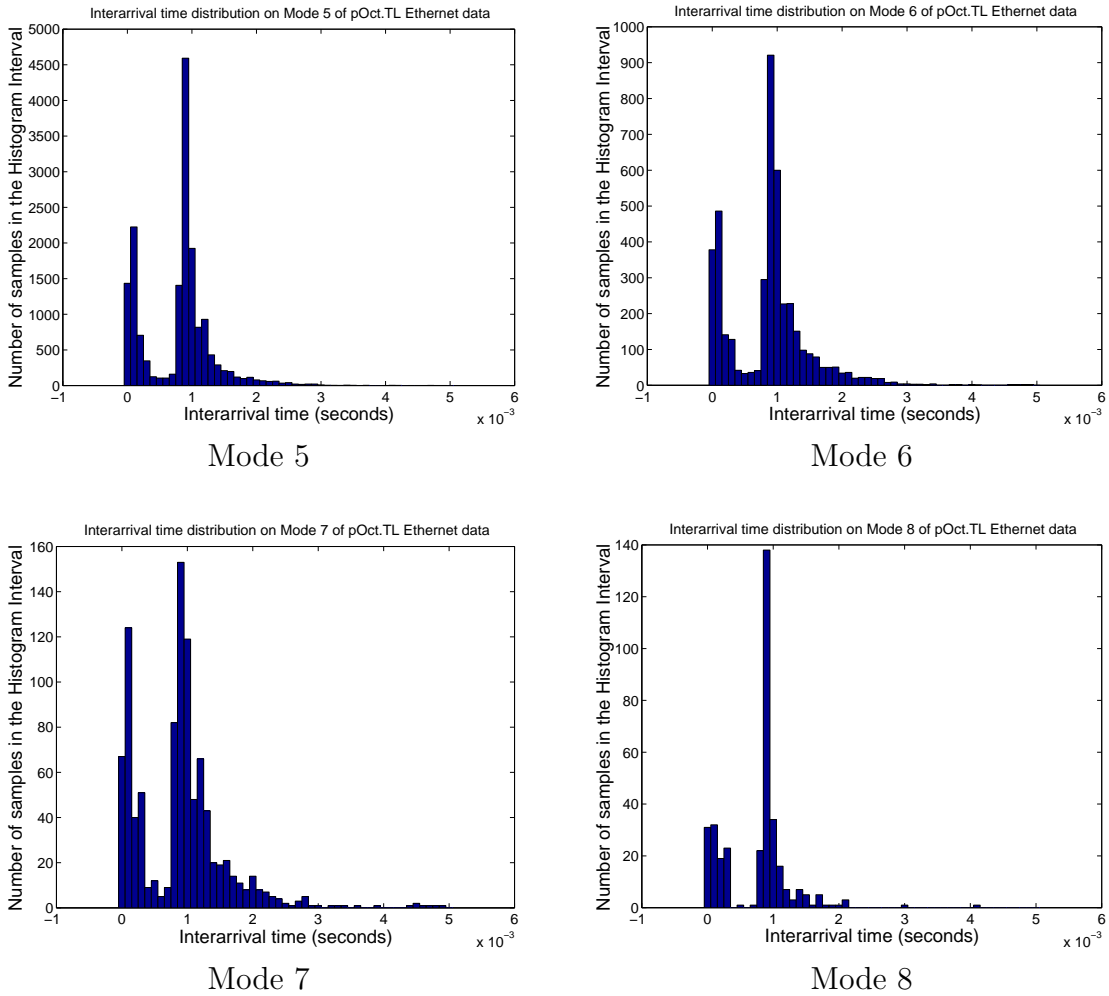
Figure 3.9: Interarrival time distributions-Modes 5-8 of pOct.TL trace

tion of exponentials) distribution. The burst length durations indicate the average amount of time a mode contributes to the traffic generation process before switching to the base mode. Clearly arrivals due to the base mode has dominance in the traffic; the base mode occupies nearly 97% of the total duration of the traffic in both the October and August traces.

3.7.3 Packet level analysis

The packet level discussion is being carried out to provide insight into the arrival mechanism within modes. The Bellcore trace contains data on packet size for every arrival which can be used in studying the characteristics of modes. Using
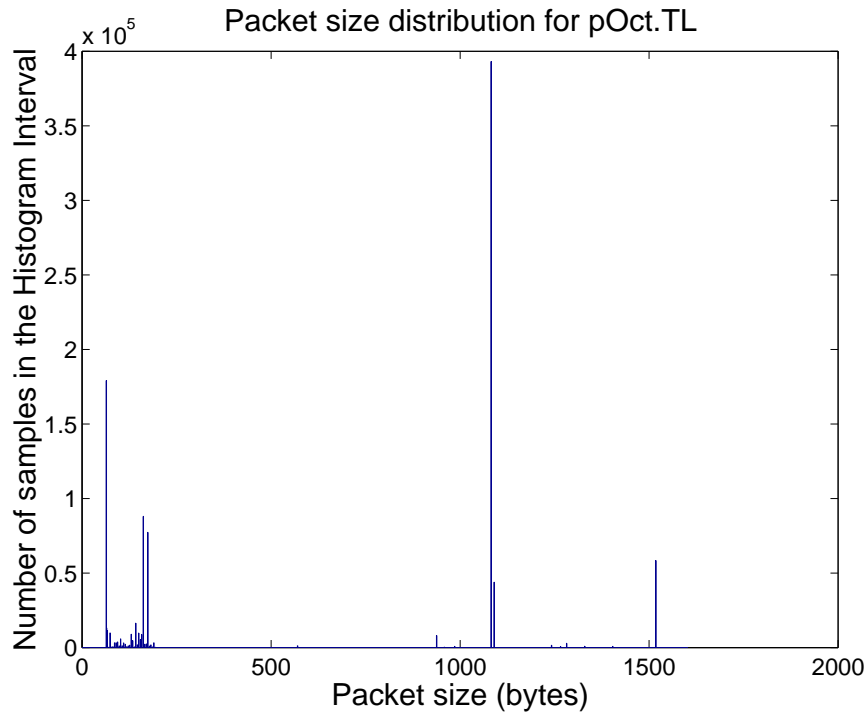
Figure 3.10: Packet size distribution-pOct.TL trace

this information, the packet size distribution of the trace and different modes can be obtained from the mode extraction process.

The packet size distribution of the pOct.TL trace (Figure 3.10) suggests that it is bimodal in nature. The smallest packet size observed was 64 bytes which corresponds to the acknowledgement (ACK) packets of Transmission Control Protocol (TCP) and the maximum packet size was 1518 bytes. The plot also gives an

Mode 1

Mode 2

Mode 3

Mode 4
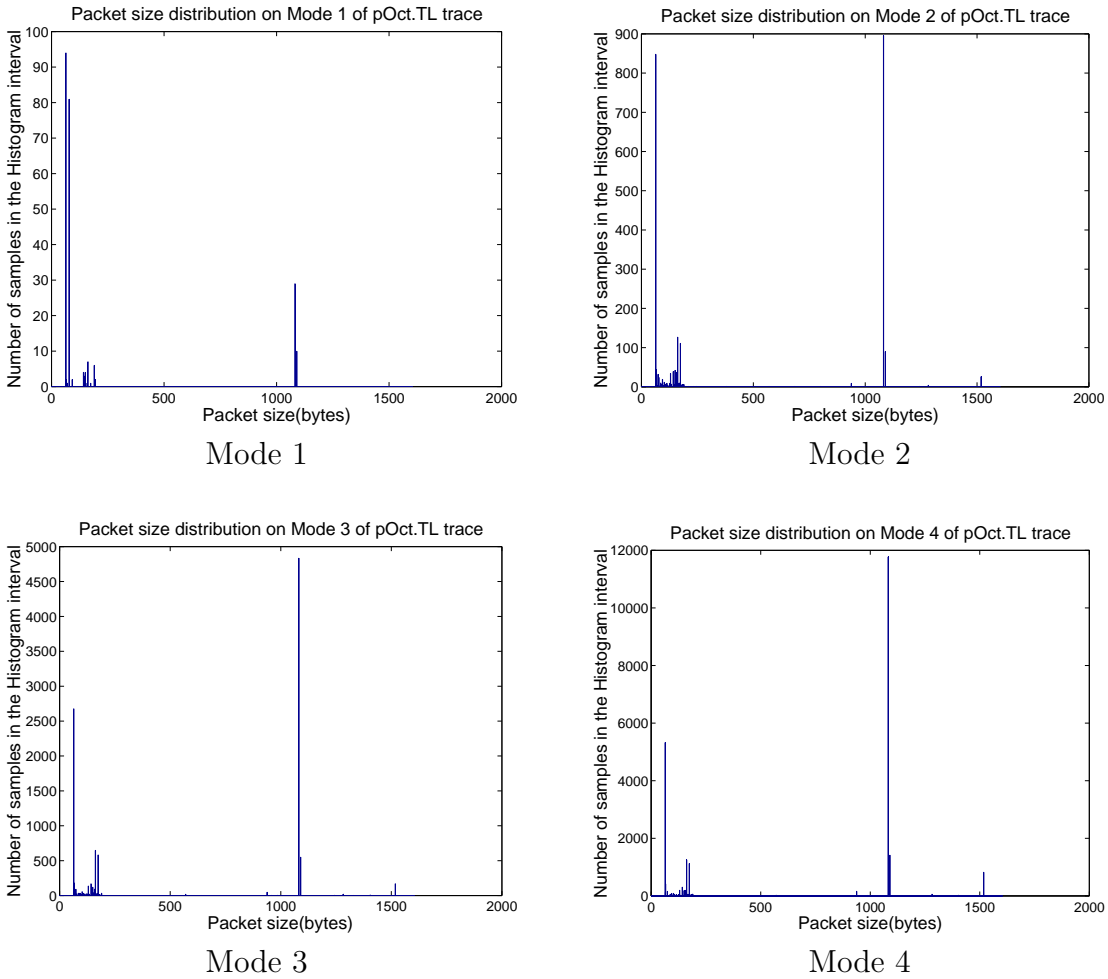
Figure 3.11: Packet size distribution-Modes 1-4 of pOct.TL trace

idea of traffic composition, with the 64 byte and 1082 byte packets clearly domi-
nating the traffic flow. A peek into the packet size distributions (Figures 3.11 and
3.12) for different modes of the traffic gave an idea on the nature of the arrivals
with shorter interarrival times. In mode 1 which is the fastest mode, the arrivals
mostly constituted 64 byte ACK packets. Because of their small size the packet
transmitter consumes less time to transmit these packets and hence they have
shorter interarrival times. Data analysis revealed that the 0.1 ms peak observed
in Figure 3.6 corresponded to the interarrival times of the ACK packets and other

Mode 5

Mode 6

Mode 7

Mode 8
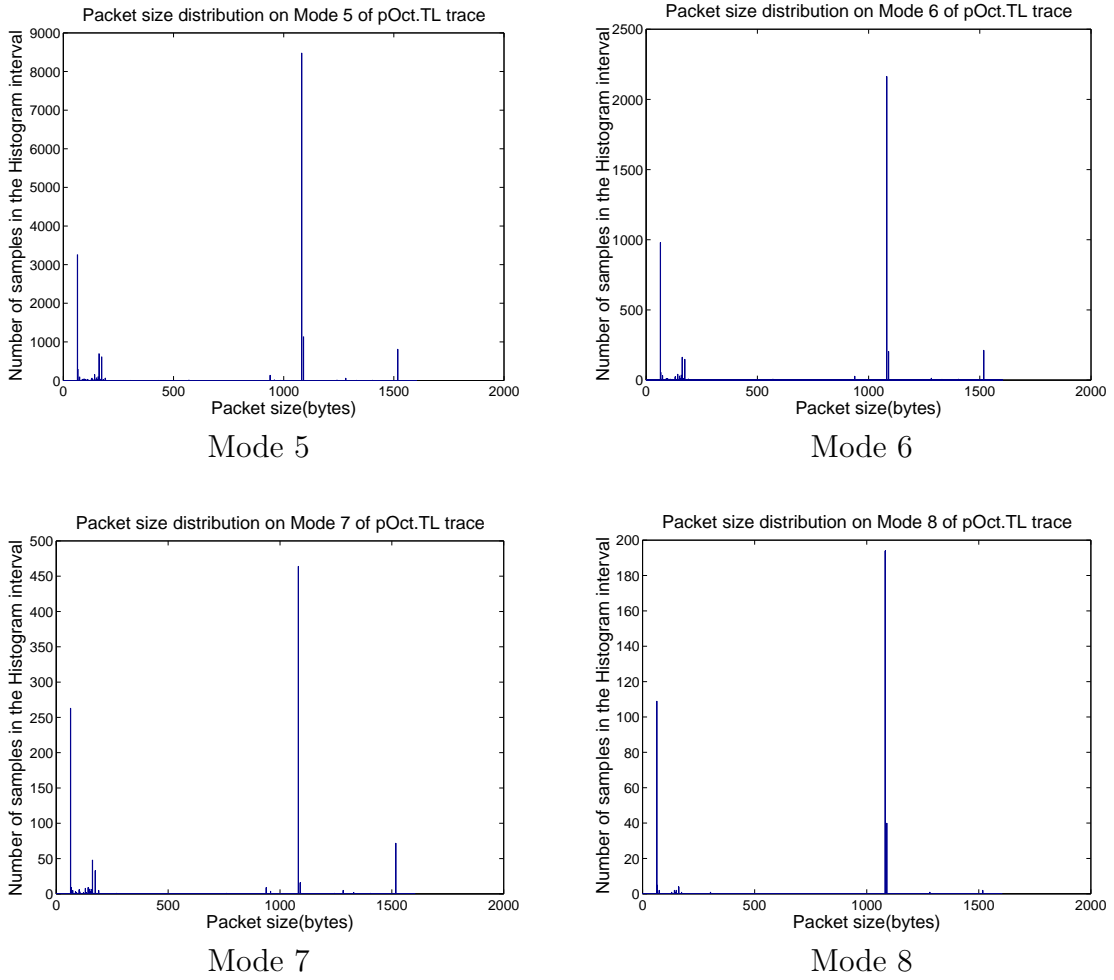
Figure 3.12: Packet size distribution-Modes 5-8 of pOct.TL trace

small sized packets. Also the 1 ms peak corresponded to the 1082 byte packets.

Plots representing other modes show a mixture of 64 byte and 1082 byte packets (compare corresponding interarrival time distribution plots). The variation in mean arrival rate of these modes is due to the different proportions in which these packets arrive in the respective modes. Very rarely do the 1518 byte packets make their presence felt in the these modes and hence are a part of the slower base mode.

A section of the traffic in the mode 1 looked like *1082 1082 162 1082 ... 1082 64 78 64... 64 1082 1082...*(the numbers representing packet size in bytes) which can be understood as a pattern of packet transmissions followed by acknowledgements. Similar patterns were also observed in other modes. This clustering pattern of the faster modes leads to rapid queue build up which has been experimentally observed in the mode extraction process. But the base mode had a well mixed pattern, clusters of packets having same size being rare and hence the queue does not build up rapidly when in base mode. Service time distribution depends on the packet size distribution but this work deals with deterministic and exponential services. If necessary, the dependence of service time distribution on mode can be incorporated in future work.

3.7.4   Correlation properties

In this section the correlation structure of the interarrival times of the traffic and of the different modes constituting the traffic are studied.

Figures 3.13 and 3.14 depict the slowly decaying autocorrelation structures of the interarrival times of the pOct.TL trace and the base mode respectively. The autocorrelation structures of the different modes in pOct.TL trace are shown in Figures 3.15 and 3.16. Modes 1, 4, 5 and 6 have pretty strong lag-1 correlation. Mode 1 has short term correlations up to a lag of 10. Mostly the autocorrelations of the modes exhibit oscillatory behavior around zero and one can assume negligible
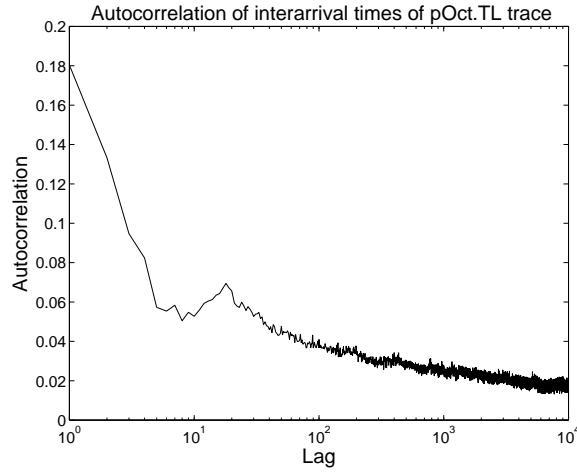
Figure 3.13: Lag-k autocorrelations-pOct.TL trace
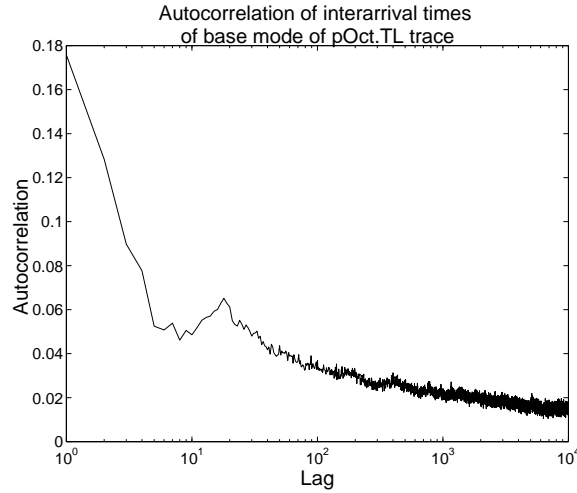
Figure 3.14: Lag-k autocorrelations-Base mode of pOct.TL trace

correlations in interarrival times of modes while fitting with analytical model. Contrary to this, the base mode has a similar correlation structure as that of the original trace, with a small drop in the autocorrelation values. Mitchell's gamma model can be used to incorporate correlations in the base mode up to a finite lag as described in section 2.5.3. Several modeling attempts discussed in chapter 2 were based on matching the entire autocorrelation structure with that of their model. However this work does not attempt at fitting the entire autocorrelation

structure.

Note: In modes 1 and 8, the number of samples is less than the lag range over which the autocorrelation is evaluated. This results in null values represented as straight line along zero, after certain lag (Figures 3.15 and 3.16).

Mode 1

Mode 2

Mode 3

Mode 4
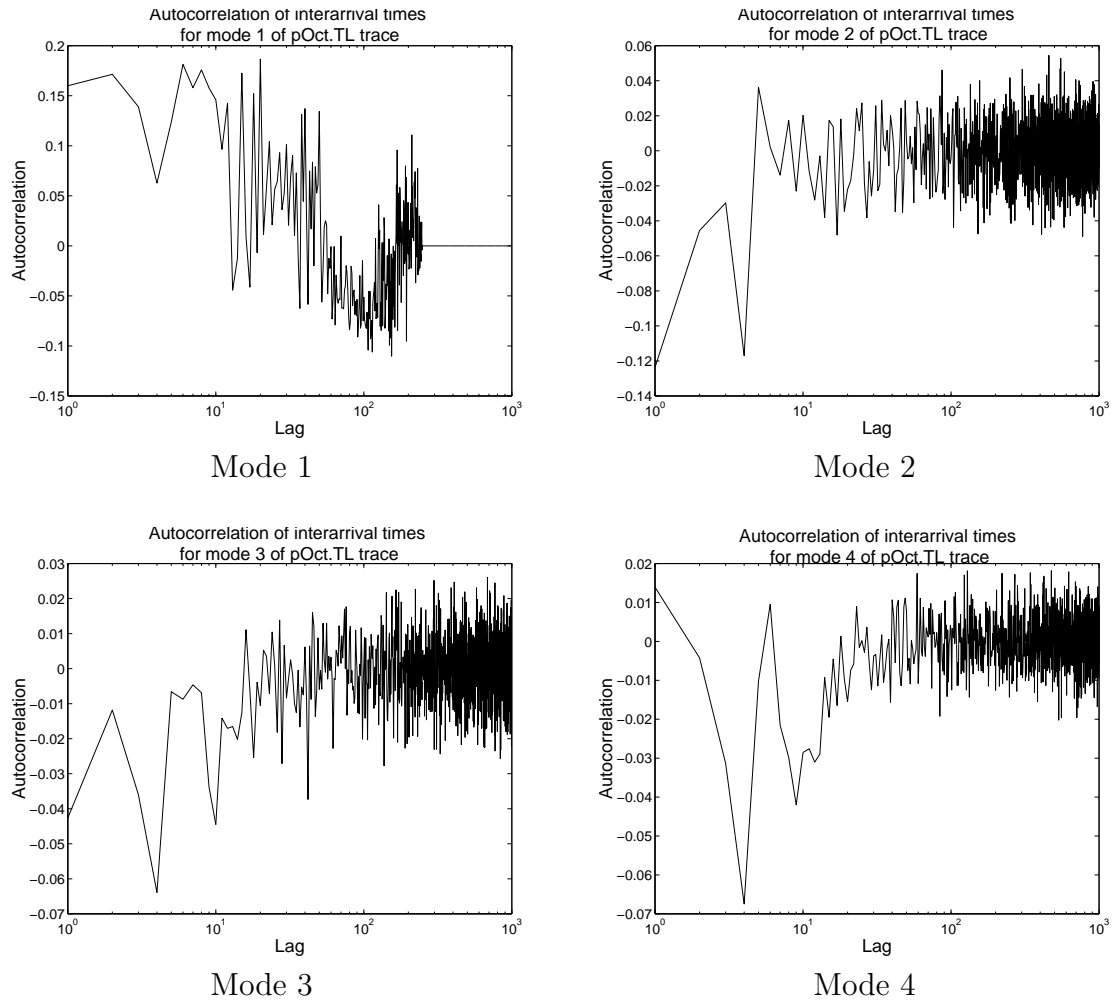
Figure 3.15: Lag-k autocorrelations-Modes 1-4 of pOct.TL trace

Mode 5

Mode 6

Mode 7

Mode 8
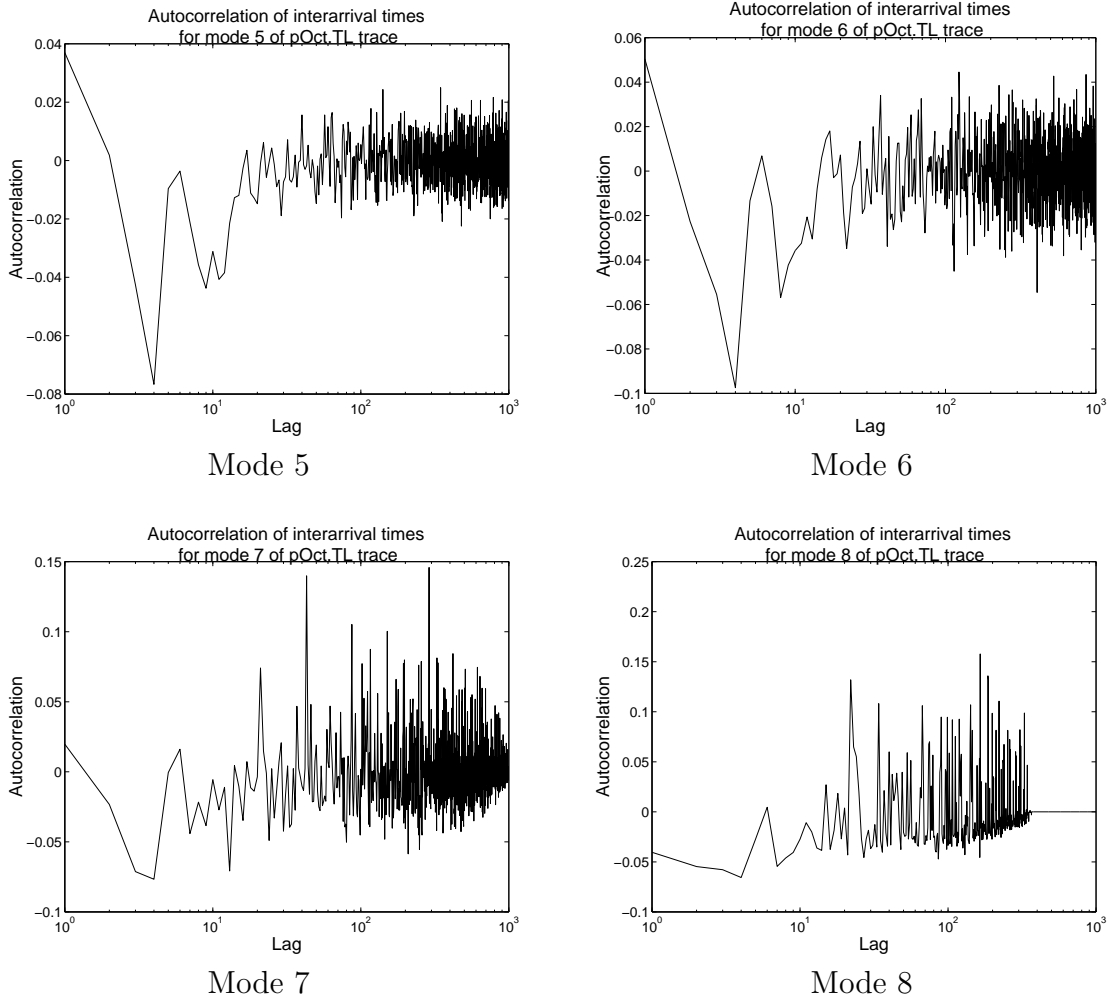
Figure 3.16: Lag-k autocorrelations-Modes 5-8 of pOct.TL trace

3.8   Summary

This chapter discussed various methods to identify and extract modes from bursty network traffic. The most efficient of the suggested methods was chosen to perform extraction. The extracted modes were classified based on the packet interarrival times within the modes. The mode which had the largest mean interarrival time was identified as the base mode which keeps the queue stable at all server utilizations. The properties of modes like mean interarrival time, lag-k correlation, burst

duration, packet size distribution were discussed and analyzed using appropriate graphs. Some of these properties will be used to develop an analytical model based on the linear algebraic queueing theory techniques discussed in Chapter 2. The discussions provided insight into the arrival pattern of different modes in the traffic. We have discussed the properties of the arrival stream in this chapter. In the next chapter, we will study the effects of the multi-modal traffic arrival stream on the queueing system by building a mathematical model based on the properties of the extracted modes.

Chapter 4

Analysis and Results

4.1   Introduction

A matrix exponential multi-modal queueing model can be constructed to match the observed characteristics of the traffic model extracted in Chapter 4. The model represents a weakly stable queueing system. This is because, over a range of utilizations, one or more of the non-base modal components of the model have mean arrival rate greater than the mean service rate, though the mean service rate is greater than the mean arrival rate of the entire traffic. The buffer and service structures can be modeled to allow comparison of its performance measures with other published modeling approaches.

4.2 $ME - Modal/ME/1/N$ model

This section deals with development of a generalized analytical model using matrix exponential multi-modal arrivals and matrix exponential service distributions with a finite buffer.

4.2.1 Spaces

The queueing system consists of five independent spaces operating simultaneously. The spaces are

1. Arrival space - the space that characterizes the interarrival times within each mode.

2. Mode space - the space that characterizes the set of modes and transitions between them.

3. Burst duration space - the space that characterizes the duration of the intervals between mode changes.

4. Queue space - the space characterizing the buffer

5. Service space - the space that characterizes the service time distribution

The arrival, mode and burst duration space combined is referred to as the traffic arrival space.

4.2.2 Arrival space

Based on the properties of modes observed, the arrivals can be modeled using a suitable distribution. The mean and coefficient of variation of the interarrival times, and correlation properties of different modes are used to construct a marginal distribution and correlation having desired properties. As discussed earlier, a simple hyperexponential distribution is used to model the mode interarrival times. A C^2 of 1.5 is assumed for modes with $C^2 < 1$.

Hyperexponential distribution

A two stage hyperexponential distribution with mean $E[X] = 1/\lambda$ and $C^2 > 1$ can be obtained as follows,

$$p_1 = \frac{1}{2}(1 - \sqrt{[(C^2 - 1)/(C^2 + 1)]}) \tag{4.1}$$

$$p_2 = 1 - p_1 \tag{4.2}$$

The starting vector for the process is $\mathbf{p_a} = [p_1 \quad p_2]$ and progress rate matrix $\mathbf{B_a}$ is given by,

$$\mathbf{B_a} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{4.3}$$

where $\lambda_1 = 2p_1\lambda$ and $\lambda_2 = 2p_2\lambda$. The summing vector is $\epsilon_a' = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Thus a matrix exponential representation $< \mathbf{p_a}, \mathbf{B_a}, \epsilon_a' >$ of the hyperexponential distribution can be obtained. Using the above technique, the interarrival time distributions for different modes can be obtained based on the assumptions made. Correlations can be introduced in the model as described in 2.4.4. The dimension of the arrival space is 2×2. Since there are nine modes, the arrival space is represented as a function of mode, m, given by $< \mathbf{p_a}(m), \mathbf{B_a}(m), \epsilon_a(m)' >$, for $m = 1, 2, \cdots, 9$.

4.2.3 Duration space

The duration space represents the burst durations in a mode. The mode duration space is represented by the pair $< \mathbf{B_d}(m), \mathbf{M_d}(m), \mathbf{L_d}(m) >$. $\mathbf{L_d}(m)$ represents the event transition matrix for the transition from mode m to base mode and $\mathbf{M_d}(m)$ represents the event transition matrix for the the transition from base mode to mode m. The matrices become scalars if exponentially distributed durations are assumed.

4.2.4 Mode space and construction of $\widehat{\mathbf{B}}_{\mathbf{mad}}$

Totally nine modes were observed in the traffic traces and hence the dimension of the mode space is nine. The mode space, arrival space and duration space are combined using the operations described below.

Kronecker products and hat spaces can be used to combine the different spaces. The subscripts identify the space in which the identified process operates and the hat represents the extension of the process into the space inferred by the equation. This is applicable throughout the following discussion. $\widehat{\mathbf{B}}_{\mathbf{a}}$, $\widehat{\mathbf{B}}_{\mathbf{d}}$ and $\widehat{\mathbf{L}}_{\mathbf{d}}$ can be written in *ad*-space as

$$\widehat{\mathbf{B}}_{\mathbf{a}}(m) = \mathbf{B}_{\mathbf{a}}(m) \otimes \mathbf{I}_{\mathbf{d}} \tag{4.4}$$

$$\widehat{\mathbf{L}}_{\mathbf{a}}(m) = \mathbf{L}_{\mathbf{a}}(m) \otimes \mathbf{I}_{\mathbf{d}} \tag{4.5}$$

$$\widehat{\mathbf{B}}_{\mathbf{d}}(m) = \mathbf{I}_{\mathbf{a}} \otimes \mathbf{B}_{\mathbf{d}}(m) \tag{4.6}$$

$$\widehat{\mathbf{L}}_{\mathbf{d}}(m) = \mathbf{I}_{\mathbf{a}} \otimes \mathbf{L}_{\mathbf{d}}(m) \tag{4.7}$$

for $m = 1, 2, \cdots, 9$.

The progress rate matrix representing the composite arrival stream $\mathbf{B}_{\mathbf{mad}}$ is obtained by summing $\widehat{\mathbf{B}}_{\mathbf{a}}$ and $\widehat{\mathbf{B}}_{\mathbf{d}}$ and then incorporating it in the mode space by a "composition" operation. In the *mad*-space it is given by,

$$
\mathbf{B_{mad}} =
\begin{bmatrix}
\widehat{\mathbf{B}}_{\mathbf{a}}(1) + \widehat{\mathbf{B}}_{\mathbf{d}}(1) & 0 & \cdots & -\widehat{\mathbf{L}}_{\mathbf{d}}(1) \\
0 & \widehat{\mathbf{B}}_{\mathbf{a}}(2) + \widehat{\mathbf{B}}_{\mathbf{d}}(2) & \cdots & -\widehat{\mathbf{L}}_{\mathbf{d}}(2) \\
\vdots & \ddots & \ddots & \vdots \\
-\widehat{\mathbf{M}}_{\mathbf{d}}(1) & -\widehat{\mathbf{M}}_{\mathbf{d}}(2) & \cdots & \widehat{\mathbf{B}}_{\mathbf{a}}(9) + \widehat{\mathbf{B}}_{\mathbf{d}}(9)
\end{bmatrix}
\tag{4.8}
$$

and the event transition matrix for the composite arrival stream in the *mad*-space is

$$
\mathbf{L_{mad}} =
\begin{bmatrix}
\widehat{\mathbf{L}}_{\mathbf{a}}(1) & 0 & \cdots & 0 \\
0 & \widehat{\mathbf{L}}_{\mathbf{a}}(2) & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \widehat{\mathbf{L}}_{\mathbf{a}}(9)
\end{bmatrix}
\tag{4.9}
$$

4.2.5 Queue-server space - 'qs' space

The queue space describes the queue and events causing an increase or decrease in the queue occupancy. The service space describes the service time distribution completely. The dimension of the queue space is the size of the buffer. The matrices $\mathbf{E_{qs}}, \mathbf{B_{qs}}, \mathbf{L_{qs}}, \mathbf{B_s}, \mathbf{L_s}$ describe the queue and service space. $\mathbf{B_s}$ is the progress

rate matrix and $\mathbf{L_s}$ is the event rate matrix of the server. The entry matrix $\mathbf{E_q}$ of size $N \times N$ represents the arrivals into the queue, where N is the buffer size. It has upper diagonal elements signifying an arrival incrementing the queue occupancy by 1. At the upper boundary, an arrival results in no change.

$$\mathbf{E_q} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \tag{4.10}$$

Extending this into the qs-space gives,

$$\mathbf{E_{qs}} = \widehat{\mathbf{E}_q} \tag{4.11}$$

The progress rate matrix in the queue-server space, $\mathbf{B_{qs}}(N \times N)$, is defined as

$$\mathbf{B_{qs}} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & \mathbf{B_s} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{B_s} \end{bmatrix} \tag{4.12}$$

The event transition matrix in the queue-server space $\mathbf{L_{qs}}(N \times N)$ is defined as

64

$$\mathbf{L_{qs}} = \begin{bmatrix} 0 & \dots & 0 & 0 \\ \mathbf{L_s} & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \mathbf{L_s} & 0 \end{bmatrix} \tag{4.13}$$

4.2.6 Combining traffic arrival space and queue-server space

The traffic arrival space and the queue-server space can be combined using the Kronecker products. The hat notation now represents Kronecker product between the traffic arrival and the queue-server space. The following equations represent the process rate matrix and the event rate matrix for the $ME/ME/1/N$ system.

$$\mathbf{B_{madqs}} = \widehat{\mathbf{B}}_{\mathbf{mad}} + \widehat{\mathbf{B}}_{\mathbf{qs}} - \widehat{\mathbf{L}}_{\mathbf{mad}} \cdot \widehat{\mathbf{E}}_{\mathbf{qs}} \tag{4.14}$$

$$\mathbf{L_{madqs}} = \mathbf{I_{mad}} \otimes \widehat{\mathbf{L}}_{\mathbf{qs}} \tag{4.15}$$

The infinitesimal generator $\mathbf{Q_{madqs}}$ is given by

$$\mathbf{Q_{madqs}} = \mathbf{L_{madqs}} - \mathbf{B_{madqs}} \tag{4.16}$$

65

The resulting infinitesimal generator will have the following structure,

$$\mathbf{Q_{madqs}} = \begin{bmatrix} -\mathbf{B}(1) & 0 & \dots & \widehat{\mathbf{L}_\mathbf{d}}(1) \\ 0 & -\mathbf{B}(2) & \dots & \widehat{\mathbf{L}_\mathbf{d}}(2) \\ \vdots & \ddots & \ddots & \vdots \\ \widehat{\mathbf{M}_\mathbf{d}}(1) & \widehat{\mathbf{M}_\mathbf{d}}(2) & \dots & -\mathbf{B}(9) \end{bmatrix} \qquad (4.17)$$

where

$$\mathbf{B}(m) = \widehat{\mathbf{B}_\mathbf{a}}(m) + \widehat{\mathbf{B}_\mathbf{d}}(m) + \widehat{\mathbf{B}_\mathbf{qs}} - \widehat{\mathbf{L}_\mathbf{a}}(m) \cdot \widehat{\mathbf{E}_\mathbf{qs}} - \widehat{\mathbf{L}_\mathbf{qs}} \qquad (4.18)$$

in the *adqs*-space.

4.3 $H_2 - Modal/M/1/N$ **model**

To make the model simpler, the burst durations of a mode are assumed to be exponentially distributed reducing the dimensionality of the duration space to a unity. This simplification modifies $\mathbf{B_{mad}}$, the composite arrival progress rate matrix, as

$$
\mathbf{B_{mad}} = \begin{bmatrix} \mathbf{B_a}(1) + \delta_1 \mathbf{I_a}(1) & 0 & \ldots & -\delta_{19}\mathbf{I_a}(1) \\ 0 & \mathbf{B_a}(2) + \delta_2 \mathbf{I_a}(2) & \ldots & -\delta_{29}\mathbf{I_a}(2) \\ \vdots & \ddots & \ddots & \vdots \\ -\delta_{91}\mathbf{I_a}(9) & -\delta_{92}\mathbf{I_a}(9) & \ldots & \mathbf{B_a}(9) + \delta_9 \mathbf{I_a}(9) \end{bmatrix} \tag{4.19}
$$

Since this model also assumes exponentially distributed services the matrix exponential parameters in the service space are converted to exponentials, simplifying the queue-server space.

$$
\mathbf{Q_{madqs}} = \begin{bmatrix} -\mathbf{B}(1) & 0 & \ldots & 0 & \delta_{19}\mathbf{I_a}(1) \\ 0 & -\mathbf{B}(2) & \ldots & 0 & \delta_{29}\mathbf{I_a}(2) \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \delta_{91}\mathbf{I_a}(9) & \delta_{92}\mathbf{I_a}(9) & \ldots & \delta_{98}\mathbf{I_a}(9) & -\mathbf{B}(9) \end{bmatrix} \tag{4.20}
$$

where, in this case

$$
\mathbf{B}(m) = \widehat{\mathbf{B}}_\mathbf{a}(m) + \widehat{\delta_\mathbf{m} \mathbf{I}_\mathbf{a}} + \widehat{\mathbf{B}}_\mathbf{qs} - \widehat{\mathbf{L}}_\mathbf{a}(m) \cdot \widehat{\mathbf{E}}_\mathbf{qs} - \widehat{\mathbf{L}}_\mathbf{qs} \tag{4.21}
$$

and

$$
\mathbf{B_{qs}} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & \mu & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mu \end{bmatrix} \qquad \mathbf{L_{qs}} = \begin{bmatrix} 0 & \dots & 0 & 0 \\ \mu & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & \mu & 0 \end{bmatrix} \qquad (4.22)
$$

To provide insight, the matrix $\widehat{\mathbf{B}}(m)$ can be detailed as,

$$
\mathbf{B}(m) = \begin{bmatrix} \mathbf{X} & -\mathbf{Y} & 0 & \dots & 0 & 0 \\ -\mathbf{S} & \mathbf{X} + \mathbf{S} & -\mathbf{Y} & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{X} + \mathbf{S} & -\mathbf{Y} \\ 0 & 0 & 0 & \dots & -\mathbf{S} & \mathbf{X} - \mathbf{Y} + \mathbf{S} \end{bmatrix} \qquad (4.23)
$$

where $\mathbf{X} = \mathbf{B_a}(m) + \delta_m \mathbf{I_a}$, $\mathbf{Y} = \mathbf{L_a}(m)$ and $\mathbf{S} = \mu \mathbf{I_a}$.

4.3.1   Solution

The steady state vector $\boldsymbol{\pi}$ for the matrix $\mathbf{Q_{madqs}}$ allows the probability distribution of the number of customers in the queueing system at steady state to be identified. $\boldsymbol{\pi}$ satisfies the following equation

$$\boldsymbol{\pi} \cdot \mathbf{Q_{madqs}} = 0 \qquad (4.24)$$

The solution can be obtained by embedding a discrete-time Markov chain at mode change instants. The mode starting probability vectors are obtained, and are then used to obtain the system steady state vector. The following equations lead to the solution for $\boldsymbol{\pi}$. Let

$$\mathbf{B} = \left[\begin{array}{ccc|c} \mathbf{B}(1) & 0 & \ldots & 0 \\[2mm] 0 & \mathbf{B}(2) & \ldots & 0 \\[2mm] \vdots & \vdots & \ddots & \vdots \\ \hline 0 & 0 & \ldots & \mathbf{B}(9) \end{array} \right] \qquad (4.25)$$

$$\mathbf{L} = \left[\begin{array}{ccc|c} 0 & \ldots & 0 & \widehat{\mathbf{L}_\mathbf{d}}(1) \\[2mm] \vdots & \ddots & \vdots & \vdots \\[2mm] 0 & \ldots & 0 & \widehat{\mathbf{L}_\mathbf{d}}(8) \\ \hline \widehat{\mathbf{M}_\mathbf{d}}(1) & \ldots & \widehat{\mathbf{M}_\mathbf{d}}(8) & 0 \end{array} \right] \qquad (4.26)$$

The above matrices are partitioned as indicated to a 2×2 form to make the solution more intuitive,

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}^{(a)} & 0 \\ 0 & \mathbf{B}^{(b)} \end{bmatrix} \qquad \mathbf{L} = \begin{bmatrix} 0 & \mathbf{L}^{(a)} \\ \mathbf{L}^{(b)} & 0 \end{bmatrix} \tag{4.27}$$

We have $\mathbf{Y} = \mathbf{B}^{-1}\mathbf{L}$ or $\mathbf{Y} = \mathbf{V}\mathbf{L}$ and hence,

$$\mathbf{Y} = \begin{bmatrix} 0 & \mathbf{V}^{(a)}\mathbf{L}^{(a)} \\ \mathbf{V}^{(b)}\mathbf{L}^{(b)} & 0 \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{Y}^{(\mathbf{a})} \\ \mathbf{Y}^{(\mathbf{b})} & 0 \end{bmatrix} \tag{4.28}$$

where

$$\mathbf{Y}^{(\mathbf{a})} = \begin{bmatrix} \mathbf{Y}^{(\mathbf{a})}(1) \\ \vdots \\ \mathbf{Y}^{(\mathbf{a})}(8) \end{bmatrix} \qquad \mathbf{Y}^{(\mathbf{b})} = \begin{bmatrix} \mathbf{Y}^{(\mathbf{b})}(1) & \cdots & \mathbf{Y}^{(\mathbf{b})}(8) \end{bmatrix} \tag{4.29}$$

and

$$\mathbf{Y}^{(\mathbf{a})}(m) = \mathbf{B}(m)^{-1} \cdot \widehat{\mathbf{L}_{\mathbf{d}}}(m) \tag{4.30}$$

$$\mathbf{Y}^{(\mathbf{b})}(m) = \mathbf{B}(m)^{-1} \cdot \widehat{\mathbf{M}_{\mathbf{d}}}(m) \tag{4.31}$$

Now define,

$$\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{Y} \tag{4.32}$$

$$\boldsymbol{\nu}^{(b)} = \boldsymbol{\nu}^{(b)} \cdot \mathbf{Y}^{(\mathbf{b})} \cdot \mathbf{Y}^{(\mathbf{a})} \tag{4.33}$$

$$\boldsymbol{\nu}^{(a)} = \boldsymbol{\nu}^{(b)} \cdot \mathbf{Y}^{(\mathbf{b})} \tag{4.34}$$

where $\boldsymbol{\nu}^{(m)}$ represents the system starting probability vector embedded at mode change instants. $\boldsymbol{\nu}^{(a)} = [\boldsymbol{\nu}_1^{(a)}\ \boldsymbol{\nu}_2^{(a)}\ \ldots\ \boldsymbol{\nu}_8^{(a)}]$ represents the starting probability vector embedded at mode change event from any other mode to the base mode and $\boldsymbol{\nu}_m^{(b)}$ represents the starting probability vector embedded at mode change event from base mode to modes 1 to 8. The vector $\boldsymbol{\nu}^{(b)}$ can be easily obtained by finding the steady state vector for the matrix $\mathbf{Y}^{(\mathbf{b})} \cdot \mathbf{Y}^{(\mathbf{a})}$ which is stochastic. From Appendix A we have,

$$[\boldsymbol{\pi}_a\quad \boldsymbol{\pi}_b] = \Big[\frac{\boldsymbol{\nu}^{(a)}\mathbf{V}^{(a)}}{\boldsymbol{\nu}^{(a)}\mathbf{V}^{(a)}\boldsymbol{\epsilon}_a' + \boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}\boldsymbol{\epsilon}_b'}\quad \frac{\boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}}{\boldsymbol{\nu}^{(a)}\mathbf{V}^{(a)}\boldsymbol{\epsilon}_a' + \boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}\boldsymbol{\epsilon}_b'}\Big] \tag{4.35}$$

The probability vectors $\boldsymbol{\pi}_a$ and $\boldsymbol{\pi}_b$ are steady-state solutions to the system being in the respective mode. The steady state vector of the entire system is given by,

$$\boldsymbol{\pi} = [\boldsymbol{\pi}_a\quad \boldsymbol{\pi}_b] \tag{4.36}$$

4.3.2 Loss probabilities

As a performance measure, packet loss probabilities associated with the system are calculated and compared with the packet loss probabilities obtained from trace driven simulation employing exponential service. Having obtained the system steady state vector, the loss probability is given by,

$$\boldsymbol{\pi}_{PLP} = \frac{\boldsymbol{\pi}(N)\mathbf{L_a}\boldsymbol{\epsilon}'}{\sum_{i=1}^{N}\boldsymbol{\pi}(i)\mathbf{L_a}\boldsymbol{\epsilon}'} \tag{4.37}$$

4.4 Results

In this section the results of the analytical model will be presented. First, the solution techniques developed for the analytical model will be verified by comparing the packet loss probability estimates of an artificially generated trace with those of the analytical model (the analytical model was built using an exponential server, so the comparison made is based on exponential service). A good match of the results will confirm that the solution techniques are correct. Then, the faithfulness of the multi-modal arrival model to the original Bellcore traffic traces will be verified by comparing their packet loss probability estimates. The multi-modal arrival model is expected to perform well at lower utilizations because the shorter or the faster time scales are modeled very well by the unstable modes.

The results and further discussions will explain the strengths and weaknesses of the suggested multi-modal arrival model in modeling the original trace. Then, the multi-modal arrival model will be compared with that of Markov modulated Poisson process (MMPP) modeling approach suggested by Anderson [1], discussing the major differences between the two approaches.

4.4.1 Verification of solution technique

In this section, the solution techniques developed will be verified, without considering efficiency of the technique. We made a conscious choice to not focus on efficient computation. Our principal purpose was to confirm (or deny) that the approach was feasible and had potential for development into an effective tool for understanding traffic.

The analytical model generated assumes a queue with modal hyperexponential arrivals, exponential burst durations and exponential service. The basic observed mode properties provided in tables 3.1 and 3.2 were used in constructing the model. Using the mathematics developed in the section 4.3, packet loss probabilities can be estimated as a means of performance measure. The loss probabilities are then compared with the loss probabilities obtained from trace driven simulations using an artificial trace (generated based on the multi-modal arrival model) as input, so as to demonstrate the correctness of the solution techniques.

The analytical model was tested for both the October trace and the August trace. The models were tested for buffer sizes up to 50 and utilizations of 40%, 50%, 60% and 70%. As the solution techniques developed were not meant to be efficient, computations become more complex and time consuming as buffer sizes increase. So tests were performed only up to buffer size of 50. Figure 4.1 shows the packet loss probabilities for different utilizations for the analytical model based on the extracted information from pOct.TL (October trace). The graphs show a good match between the analytical results and the simulation for all the buffer sizes tested but for the utilization of 60%. Similarly figure 4.2 show the comparison plots for the August trace data. The above result serves as reassurance that the solution methodologies for the analytical model are correct.

4.4.2 Modal traffic vs Original traffic - Performance comparison

In this section the faithfulness of the multi-modal arrival traffic model to the original traffic will be demonstrated. The validity of the solution techniques has already been established. But the techniques are not employed in this section because the aim of this section is to compare the performance of the multi-modal model with that of the original trace.

Using the multi-modal traffic arrival model artificial traces with 1 million arrivals were generated (based on extracted data from October and August traces)
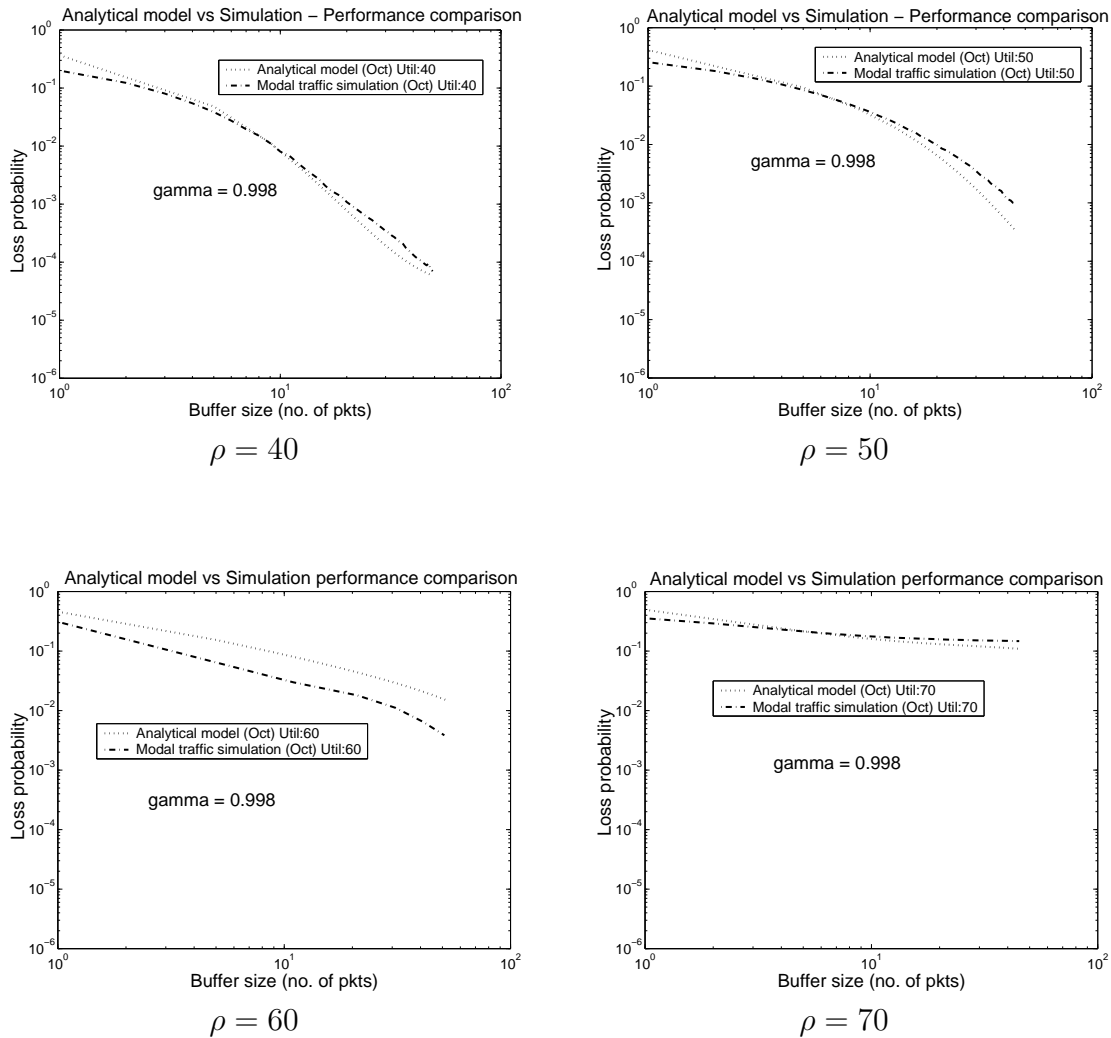
Figure 4.1: Loss probabilities for analytical model (Oct) ρ=40, 50, 60, 70

with $\gamma = 0$ for the base mode and was fed to a queue with exponential server. As discussed in Chapter 2, $\gamma = 0$ implies zero autocorrelations and γ approaching 1 implies presence of autocorrelations with long lag times. As a performance measure, the packet loss probabilities were obtained at different utilizations for both the multi-modal artificial traces and the Bellcore traces. Figure 4.3 gives the comparison of packet loss probabilities for the original October trace and the generated multi-modal traffic trace. Though autocorrelations are absent in the
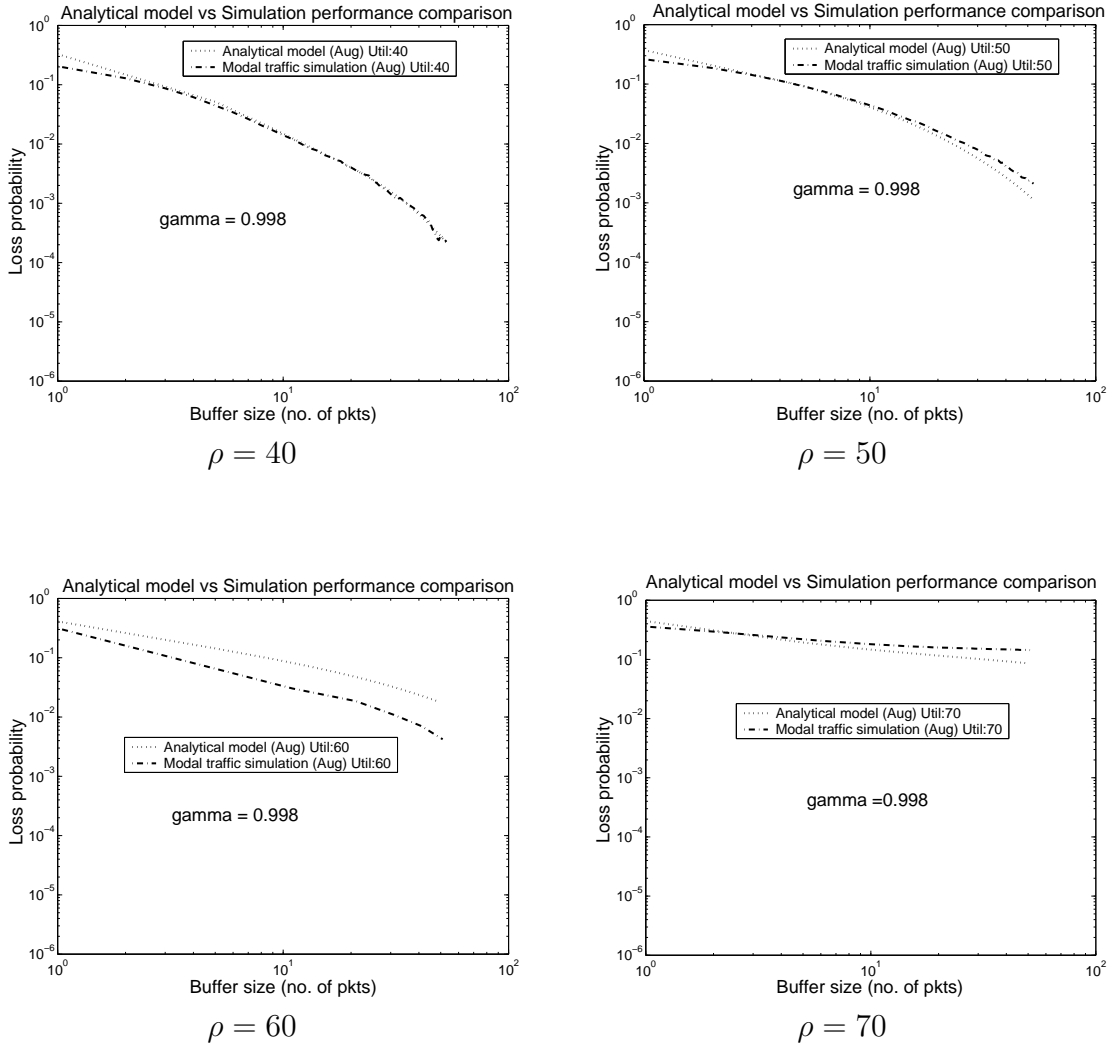
Figure 4.2: Loss probabilities for analytical model (Aug) ρ=40, 50, 60, 70

base mode ($\gamma = 0$) there is a good match at lower utilizations of 40% and 50%, but

at utilization of 60% the absence of correlations in the base mode can be felt. The

results deviate significantly at utilization of 60%. This experiment demonstrates

the importance of base mode correlation for performance predictions at higher

utilizations and the importance of unstable modes at lower utilizations. To study

the effect of autocorrelations on the loss probabilities, another sample trace was

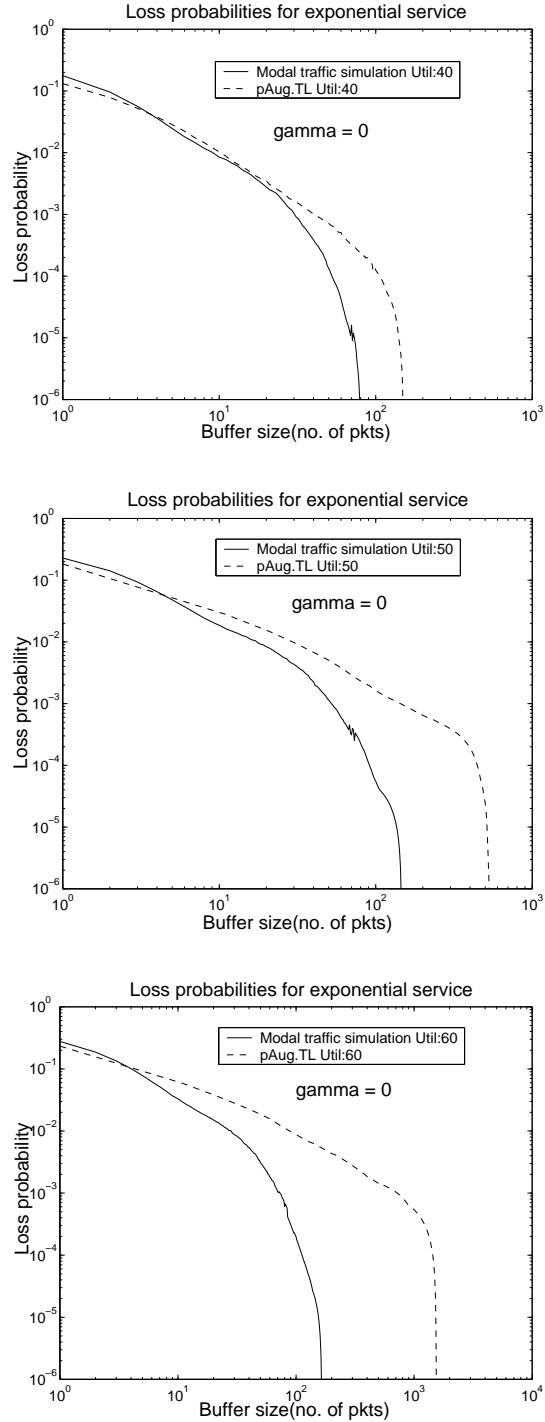generated with $\gamma = 0.998$ for the base mode. The results for this trace are shown
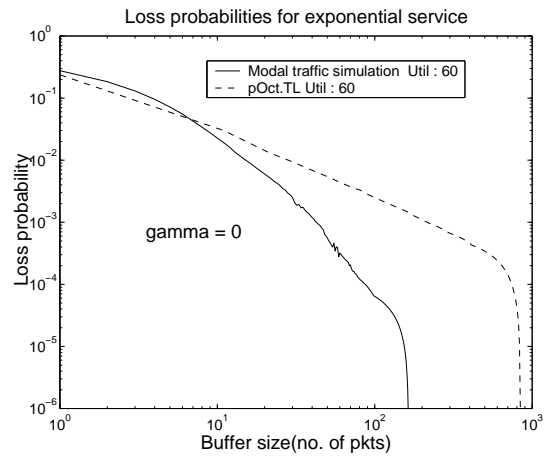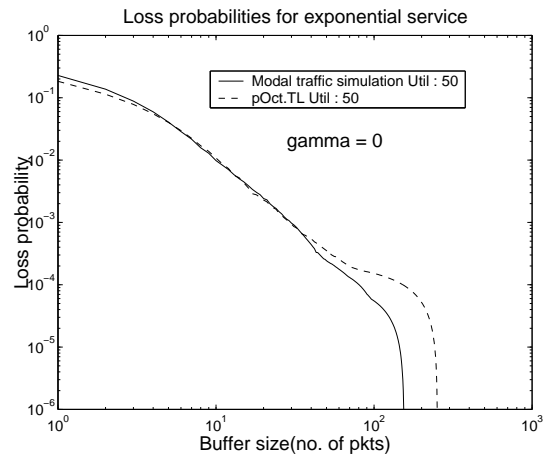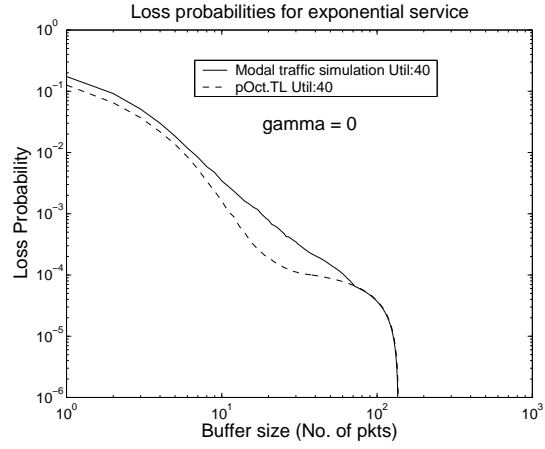
Figure 4.3: Loss probabilities for modeled modal traffic and actual traffic (Oct trace, ρ=40, 50 and 60, γ=0)

in the figure 4.4. Introduction of correlations in the base mode causes a jump in the values of loss probabilities, but there is overestimation. The gamma value chosen provides autocorrelations extending up to a lag of 2000 as shown in figure 4.5. This provides a finite lag correlation structure to the model though it doesn't fit the original correlation structure. A model that capture correlations up to a large finite lag can be good approximation for performance evaluation as discussed in Chapter 2. Comparison plots (figure 4.6 for $\gamma = 0.998$ and figure 4.7 for $\gamma = 0$) for the Bellcore August trace also illustrate the effect of correlations as discussed above. These results again suggest the importance of modeling the base mode accurately. But the results are encouraging in the sense that, in spite of the crude assumptions made in modeling the base mode, the performance estimation has been "pretty good" at lower utilizations and "fair" in the presence of correlations at higher utilizations.

4.4.3 Comparison with Anderson model

Anderson et al. [1] use a model consisting of superposition of two-state MMPP's. They fit the correlation structure of the traffic arrival process to their model over a finite range of time scales, using a fitting algorithm. They compare the tail distribution of the queue with infinite buffer and deterministic service, obtained using trace driven simulation of the data, with their analytical results. The performance results are tested for both the October and August Bellcore traces and
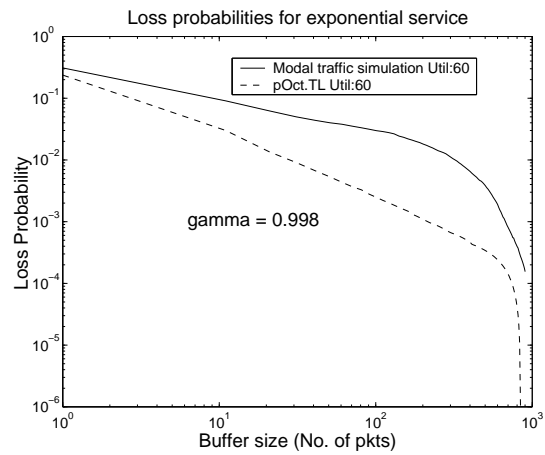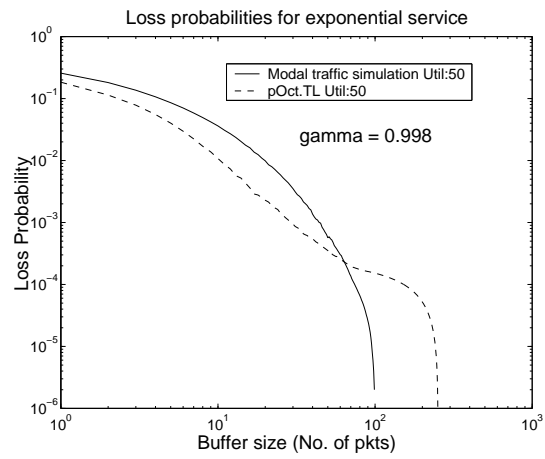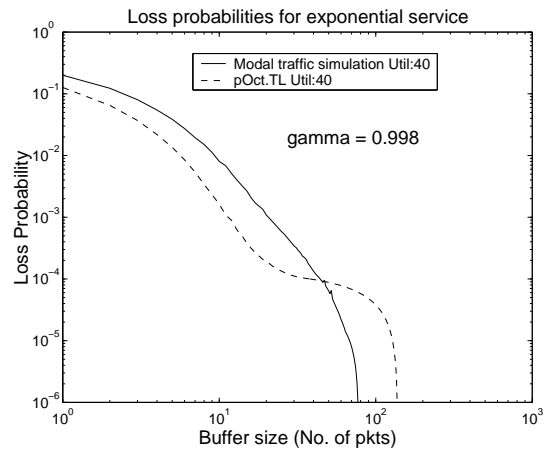
Figure 4.4: Loss probabilities for modeled modal traffic and actual traffic (Oct trace, ρ=40, 50 and 60, γ=0.998)
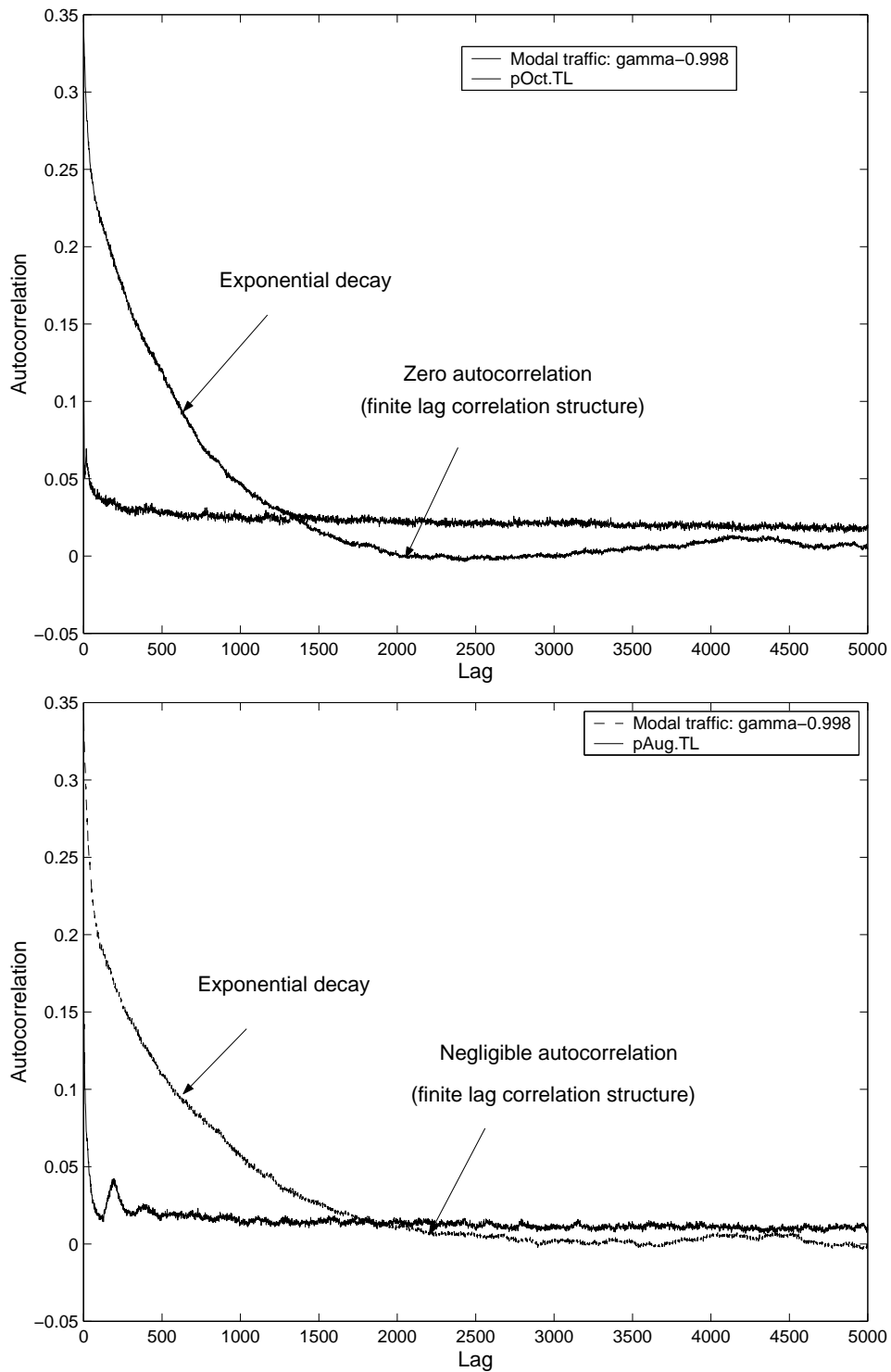
Figure 4.5: Finite lag correlation structure :October and August data

the solution techniques are based on the matrix analytical methods. Their results

deviated significantly at lower utilizations while the performance was better at
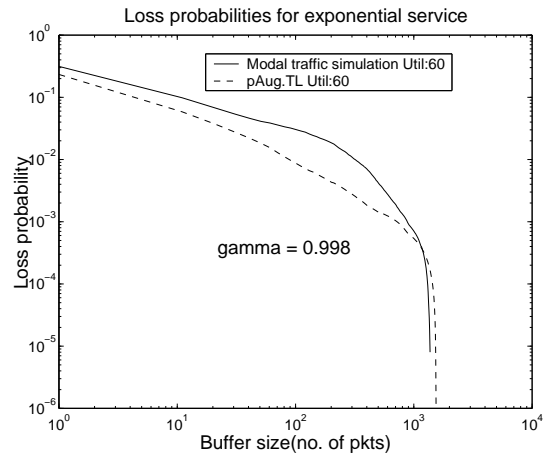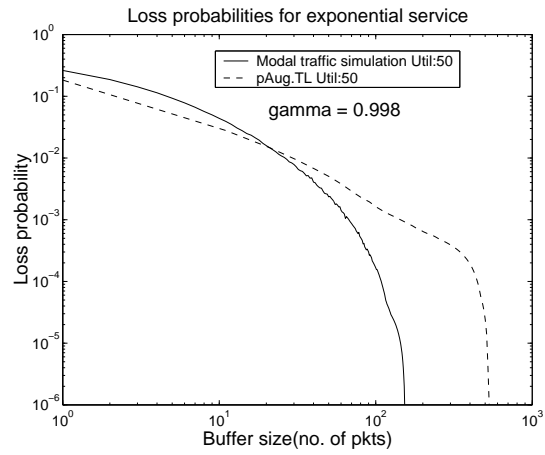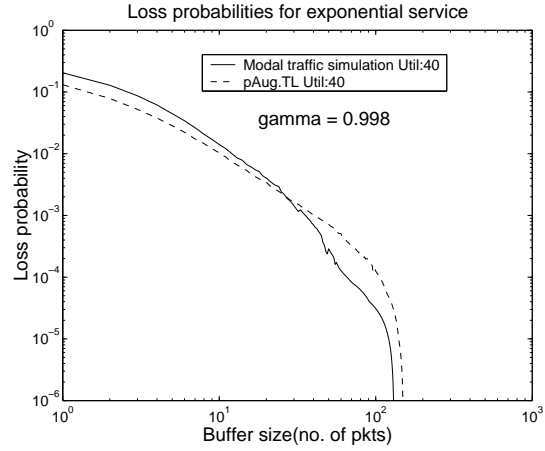
Figure 4.6: Loss probabilities for modeled modal traffic and actual traffic (Aug trace, ρ=40, 50 and 60, γ=0.998)

Figure 4.7: Loss probabilities for modeled modal traffic and actual traffic (Aug trace, ρ=40, 50 and 60, γ=0)

higher utilizations. Also, the performance results for the August trace was poor, as their model overestimated the tail probabilities. They concluded that the

queueing behavior cannot be predicted accurately by fitting the first and second order properties of the counting process alone. This section compares the performance of modal traffic modeling approach and the MMPP modeling approach of Anderson.

The artificial multi-modal traffic arrival trace generated with $\gamma = 0.998$ for the base mode was fed to an infinite queue with deterministic server. Anderson used a deterministic server and estimated the probability of queue length exceeding a certain limit for utilizations of 40%, 50% and 60%. For the October trace, figures 4.8 to 4.13 give comparisons of the results obtained by Anderson and the results of the multi-modal traffic model at different utilizations, for $\gamma = 0.998$. In the case of the multi-modal model, due to high autocorrelations at lower lags, the probabilities are overestimated, but there is a significant drop at larger queue sizes. Anderson's model does not predict the probabilities well after a buffer size of 10 at $\rho = 40\%$ but shows a good match for $\rho = 50$ and 60 up to a certain buffer size. In case of the August trace (figures 4.14 to 4.19), clearly the multi-modal model is able to predict the probabilities better than that of the Anderson model for $\rho = 40\%$ and $\rho = 60\%$, for $\gamma = 0.998$.

From the above results it is evident that Anderson's model did not model the instabilities in the traffic which is essential for performance at lower utilizations, while the multi-modal model has an advantage in this respect. The weakness of the multi-modal model is that the base mode modeling is crude, as explained earlier. This causes deviation in results at higher utilizations but again, the results

are promising considering the deficiencies in modeling the base mode.

4.4.4   Summary of experimental results

From the above results it can be inferred that the presence of autocorrelations at large lags (as observed in the extracted base mode) play an important role in influencing the queue lengths at higher utilizations. The marginal distribution and correlation structure of the base mode also play a vital role in providing close match to reality. Higher γ values (close to 1) do provide long lag correlations but they decay exponentially which is very different from the correlation structure of the base mode.

In case of exponential service with $\gamma = 0$ the packet loss probabilities were well estimated at lower utilizations, strengthening the importance of the concept of weak stability. At lower utilizations the impact of the stable or the base mode is less. The losses are mainly due to the presence of unstable modes and the short range correlations (lag < 100) that arise because of the weak leakages from faster unstable modes to the slower base mode. The short range correlations confirm that communication between modes can give rise to dependence in the arrival mechanism. It is likely that the long range correlation structure of the entire traffic stream comes from the weak communication between the slower modes within the base mode itself (the base mode could be structured!) and has to be investigated.

Chapter 5

Conclusions and Future research

The goal of this work was to show that bursty dependent network traffic can be characterized by a multi-modal cpMc possessing the property of near complete decomposability. The multi-modal traffic model provides much needed insight into the traffic arrival mechanism which other published modeling approaches failed to address.

We were successful in identifying and extracting modes from publicly available Bellcore traffic traces. Different methods were formulated to identify and extract the modes each having their own advantages and disadvantages. The most efficient method was used in the extraction process; this method did not need any manual intervention and was effectively used to obtain the properties of different modes. An analysis of the mode properties showed the nature of arrivals within each mode. The analysis illustrated that the faster modes have negligible

correlations in the interarrival times and that the base mode is characterized by correlations of interarrival times extending to long time lags very similar to the correlation structure of the real trace. We were able to split the traffic into meaningful and probabilistically interpretable components such that every component can be individually analyzed and modeled.

We were able to model the modal arrival structure using matrix exponential extensions to Markov analysis, which is remarkable. We used the hyperexponential distribution with physically interpretable phases to model the marginal structure of the interarrival times of different modes and used the Mitchell method of introducing correlations into the base mode. As a one parameter model, the Mitchell model can be expected to only crudely model the correlation structure of the base mode. This modeling approach is different from other available arrival models in the sense that the parameters used to build our model are directly obtained from the traffic, without any need for fitting algorithms used in most of the models. Linear algebraic queueing theory techniques were effectively used to model the queueing system with modal arrival inputs. We followed a modular development of the analytical model allowing deeper understanding of the model. The solution techniques based on LAQT allowed us to investigate the impact of different modes on the packet loss probability of the system and yielded deep insight into the nature of the system.

The correctness of the analytical solution technique was established by the good match of the packet loss probability estimates of the analysis with that

obtained from the artificial-trace driven simulation. Performance comparison was also made with original traces and Anderson's modeling approach, to study the effectiveness of the proposed multi-modal model. Overall, performance results showed a pretty good fit for the lower utilizations, highlighting the contribution from the multi-modal representation. This supports the presence of modes in the original traffic and suggests that a weakly stable queueing model with modal input is a good approximation for systems with bursty traffic inputs. This result also illustrates that a traffic arrival model that does not satisfy the property of *long range dependence* can predict the performance quite accurately at lower utilizations. The finite lag correlations of the base mode introduced in the form of γ provided good estimates of loss probabilities at smaller buffer sizes over varying utilizations.

The performance deviated significantly for higher buffer sizes at higher utilizations. This might be due to the lack of accuracy in modeling the base mode correlation structure. This stresses the need for a better approach to capture the correlations in the base mode.

We conjecture that the multi-modal traffic model will provide accurate performance estimates if the base mode is modeled properly. The base mode could be structured because there are no reasons to believe that the modal structure does not extend even into the base mode. It is possible that modeling the sub-modes of the base mode and their weak interactions would automatically model the correlation structure of the base mode. The extraction technique suggested is

not suitable to analyze the base mode structure and the task is out of scope for this thesis. Additional research is necessary to extract the sub-modes from the base mode and to exploit them in refining the method.

We conclude by listing some of the major contributions of this work,

1. Introduced the notion of modes in bursty network traffic.

2. Developed efficient methods to extract the modes from trace data.

3. Analyzed the characteristics of modes.

4. Developed a mathematical model based on the multi-modal arrival model using linear algebraic queueing theory techniques.

5. Evaluated the performance of the analytical model and compared it with the trace driven simulations. A comparison study with another well known modeling approach was also done.

6. Showed that a traffic arrival model with long but finite duration correlations can predict the queueing behavior accurately at lower utilizations.

Bibliography

[1] A.T. Anderson and B.F. Nielsen. A Markovian approach for modeling packet traffic. *IEEE Journal on selected areas in communications*, 16(5):719–732, 1998.

[2] P. J. Courtois. *Decomposability: Queueing and computer system applications*. Academic press, New York, 1977.

[3] D. R. Cox. *Long-Range Dependence: A Review*. The Iowa State University Press, Ames, Iowa, 1984.

[4] Mark Crovella and Azer Bestavros. Self-similarity in world wide web traffic: evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[5] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4:209–223, 1996.

[6] Do Young Eun and Ness B. Shroff. A measurement-analytic approach for QoS estimation in a network based on the dominant time scale. *IEEE/ACM Transactions on Networking*, 11(2):222–235, April 2003.

[7] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3-4):245–279, April 1997.

[8] Mark W. Garrett and Walter Willinger. Analysis, modeling and generation of self-similar VBR video traffic. In *SIGCOMM*, pages 269–280, 1994.

[9] Matthias Grossglauser and Jean-Chrysostome Bolot. On the relevance of long-range dependence in network traffic. *IEEE/ACM Transactions on Networking*, 7(5):629–640, 1999.

[10] Bruce Hajek and Linhai He. On variations of queue response for inputs with the same mean and autocorrelation function. *IEEE/ACM Transactions on Networking*, 6(5):588–598, 1998.

[11] Predrag R. Jelenkovic, Aurel A. Lazar, and Nemo Semret. The effect of multiple time scales and subexponentiality in MPEG video streams on queueing behavior. *IEEE Journal of Selected Areas in Communications*, 15(6):1052–1071, 1997.

[12] Robert Etheredge Juliano. A linear algebraic based solution method for queing systems with highly correlated arrival process. *Ph.D Thesis, The University of Kansas*, January 2000.

[13] Shoji Kasahara. Internet traffic modeling: Markovian approach to self-similar traffic and prediction of loss probability for finite queues. *IEICE Transactions on Communications*, E84-B(8):2134–2141, August 2001.

[14] Lester R. Lipsky. *Queueing Theory : A linear algebraic approach*. Macmillan Publishing Company, 1992.

[15] B. B. Mandelbrot. *The Fractal geometry of nature*. Freeman, New York, 1983.

[16] K. Mitchell and A. van de Liefvoort. Approximation models of feed-forward G/G/1/N queueing networks with correlated arrivals. *Operations research Letters*, 51(2-4):137–152, February 2003.

[17] K. Mitchell, A. van de Liefvoort, and J. Place. Correlation properties of the token leaky bucket departure process. *Computer Communications*, 21:1010–1019, March 1998.

[18] Marcel F. Neuts. *Algorithmic Probability : A Collection of Problems (Stochastic Modeling)*. Chapman and Hall, London, 1995.

[19] Kihong Park, Gitae Kim, and Mark Crovella. On the relationship between file sizes, transport protocols, and self-similar network traffic. *Proceedings of the 1996 International Conference on Network Protocols (ICNP '96)*, page 171, October 29-November 01 1996.

[20] Kihong Park and Walter Willinger. *Self-similar network traffic and performance evaluation*. Wiley, 2000.

[21] Vern Paxson and Sally Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

[22] San qi Li and Chia-Lin Hwang. Queue response to input correlation functions:continuous spectral analysis. *IEEE/ACM Transactions on Networking*, 1(6):678–692, December 1993.

[23] Vinay J. Ribeiro, Rudolf H. Riedi, Matthew S. Crouse, and Richard G. Baraniuk. Multiscale queuing analysis of long-range-dependent network traffic. *IEEE INFOCOM 2000 - The Conference on Computer Communications*, (1):1026 – 1035, March 2000.

[24] Ronn Ritke, Xiaoyan Hong, and Mario Gerla. Contradictory relationship between Hurst parameter and queueing performance (extended version). *Telecommunication Systems*, 16(1-2):159–175, January 2001.

[25] Stephan Robert and Jean-Yves Le Boudec. New models for pseudo self-similar traffic. *Performance Evaluation*, 30(1-2):57–68, 1997.

[26] B. Ryu and A. Elwalid. The importance of long range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *In Proceedings ACM SIGCOMM '96*, pages 3–14, 1996.

[27] P. Salvador, R. Valadas, and A. Pacheco. Multiscale fitting procedure using Markov modulated Poisson processes. *Telecommunication systems*, 23(1/2):123–148, June 2003.

[28] M. Taqqu, V. Teverovsky, and W. Willinger. Estimators for long–range dependence: An empirical study. *Fractals*, 3(4):785–788, 1995.

[29] A. van de Liefvoort. The moment problem for continuous distributions. Technical report, 1990.

[30] W.E.Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, Feb 1994.

[31] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson. Self-similarity through high variability: statistical analysis of ethernet LAN traffic at the source level. *In Proceedings ACM SIGCOMM 95*, pages 100–113, 1995.

[32] T. Yoshihara, S. Kasahara, and Y. Takahashi. Practical time-scale fitting of self-similar traffic with Markov modulated Poisson process. *Telecommunication Systems*, 17(1-2):185–211, 2001.

[33] R. L. Zarling. *Numerical solution of nearly decomposable queueing networks*. PhD thesis, University of North Carolina, 1976.

Appendix A

Relationship between π and ν

The steady state vector $\boldsymbol{\pi}$ satisfies the following equations,

$$\boldsymbol{\pi}\mathbf{Q} = 0 \tag{A.1}$$

$$\pi(\mathbf{L} - \mathbf{B}) = 0 \tag{A.2}$$

$$\boldsymbol{\pi}\mathbf{L} = \boldsymbol{\pi}\mathbf{B} \tag{A.3}$$

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{L}\mathbf{B}^{-1} \tag{A.4}$$

The equilibrium starting probability vector $\boldsymbol{\nu}$ satisfies the following equations,

$$\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{Y} \tag{A.5}$$

$$\boldsymbol{\nu} = \boldsymbol{\nu}\mathbf{B}^{-1}\mathbf{L} \tag{A.6}$$

$$(\boldsymbol{\nu}\mathbf{B}^{-1}) = (\boldsymbol{\nu}\mathbf{B}^{-1})\mathbf{L}\mathbf{B}^{-1} \tag{A.7}$$

From equations A.4 and A.7 we have proportionality between $\boldsymbol{\pi}$ and $\boldsymbol{\nu}\mathbf{B}^{-1}$

$$\boldsymbol{\pi} \propto \boldsymbol{\nu}\mathbf{B}^{-1} \tag{A.8}$$

Let $\alpha = [\alpha_a \quad \alpha_b]$ be the weights associated with the proportionality. Then,

$$[\boldsymbol{\pi_a} \quad \boldsymbol{\pi_b}] = [\alpha_a \boldsymbol{\nu}^{(a)} \quad \alpha_b \boldsymbol{\nu}^{(b)}]\mathbf{B}^{-1} \tag{A.9}$$

We know,

$$\boldsymbol{\pi}\mathbf{Q} = 0 \tag{A.10}$$

$$-\alpha_a \boldsymbol{\nu}^{(a)}\mathbf{V}^{(a)}\mathbf{B}^{(a)} + \alpha_b \boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}\mathbf{L}^{(b)} = 0 \tag{A.11}$$

$$-\alpha_a \boldsymbol{\nu}^{(a)} + \alpha_b \boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}\mathbf{L}^{(b)} = 0 \tag{A.12}$$

$$-\alpha_a \boldsymbol{\nu}^{(a)} + \alpha_b \boldsymbol{\nu}^{(b)}\mathbf{Y}^{(b)} = 0 \tag{A.13}$$

$$-\alpha_a \boldsymbol{\nu}^{(a)} + \alpha_b \boldsymbol{\nu}^{(a)} = 0 \tag{A.14}$$

$$\alpha_a = \alpha_b \tag{A.15}$$

Also,

$$\boldsymbol{\pi}\boldsymbol{\epsilon}' = 1 \tag{A.16}$$

$$\alpha_a \boldsymbol{\nu}^{(a)}\mathbf{V}^{(a)}\boldsymbol{\epsilon_a}' + \alpha_b \boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}\boldsymbol{\epsilon_b}' = 1 \tag{A.17}$$

$$\alpha_a \boldsymbol{\nu}^{(a)}\mathbf{V}^{(a)}\boldsymbol{\epsilon}_a' + \alpha_b \boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}\boldsymbol{\epsilon}_b' = 1 \tag{A.18}$$

$$\alpha_a(\boldsymbol{\nu}^{(a)}\mathbf{V}^{(a)}\boldsymbol{\epsilon}_a' + \boldsymbol{\nu}^{(b)}\mathbf{V}^{(b)}\boldsymbol{\epsilon}_b') = 1 \tag{A.19}$$

So,

$$\alpha_a = \alpha_b = \frac{1}{\boldsymbol{\nu}^{(a)}\mathbf{V}^{(\mathrm{a})}\boldsymbol{\epsilon}'_a + \boldsymbol{\nu}^{(b)}\mathbf{V}^{(\mathrm{b})}\boldsymbol{\epsilon}'_b} \tag{A.20}$$

Hence,

$$\boldsymbol{\pi}_a = \frac{\boldsymbol{\nu}^{(a)}\mathbf{V}^{(\mathrm{a})}}{\boldsymbol{\nu}^{(a)}\mathbf{V}^{(\mathrm{a})}\boldsymbol{\epsilon}'_a + \boldsymbol{\nu}^{(b)}\mathbf{V}^{(\mathrm{b})}\boldsymbol{\epsilon}'_b} \tag{A.21}$$

$$\boldsymbol{\pi}_b = \frac{\boldsymbol{\nu}^{(b)}\mathbf{V}^{(\mathrm{b})}}{\boldsymbol{\nu}^{(a)}\mathbf{V}^{(\mathrm{a})}\boldsymbol{\epsilon}'_a + \boldsymbol{\nu}^{(b)}\mathbf{V}^{(\mathrm{b})}\boldsymbol{\epsilon}'_b} \tag{A.22}$$

Appendix B

Mode properties : pAug.TL
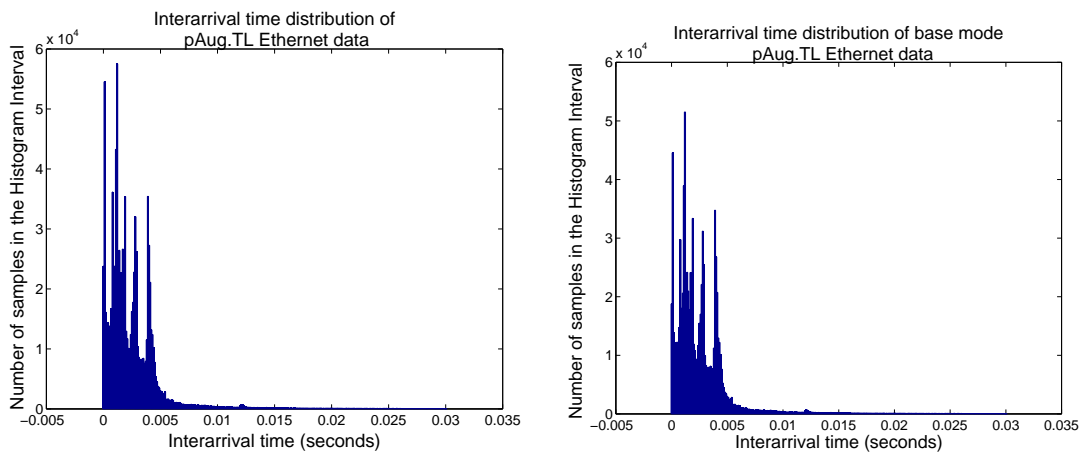
Figure B.1: Interarrival time distribution: pAug.TL and base mode
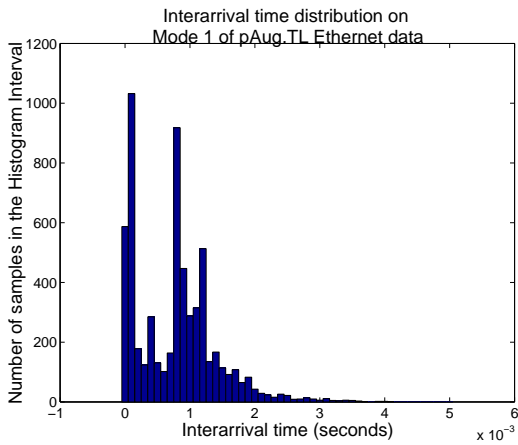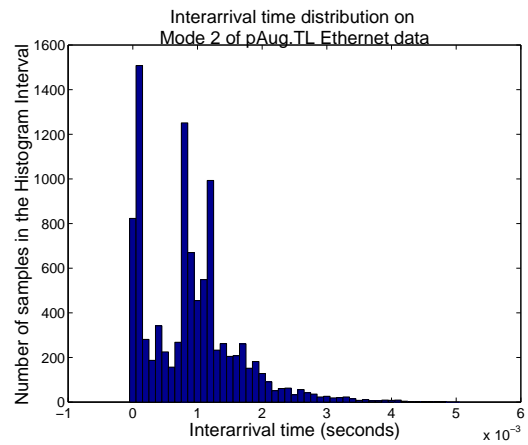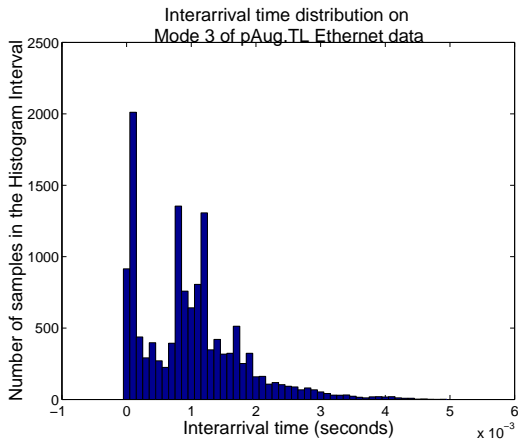
Figure B.2: Interarrival time distributions-Modes 1-4 of pAug.TL trace
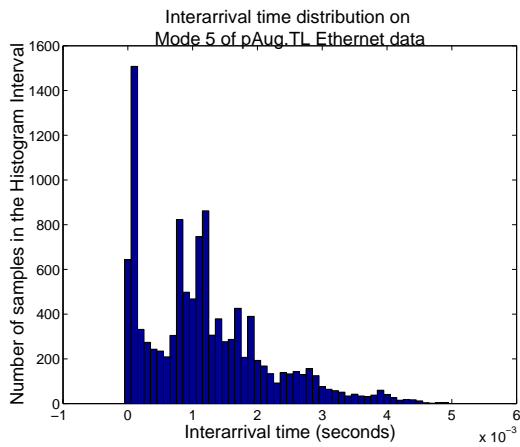
Mode 5
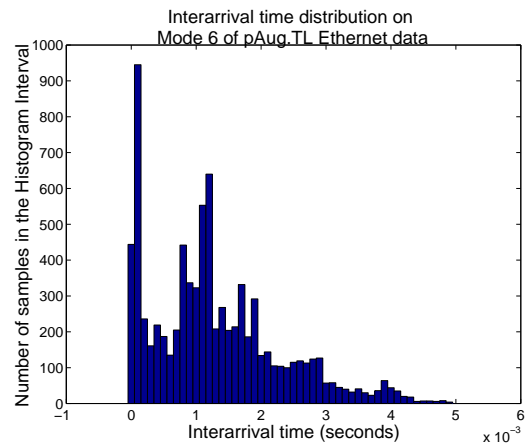
Mode 6
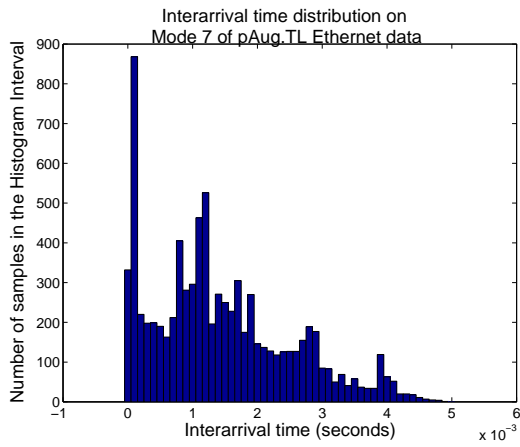
Mode 7
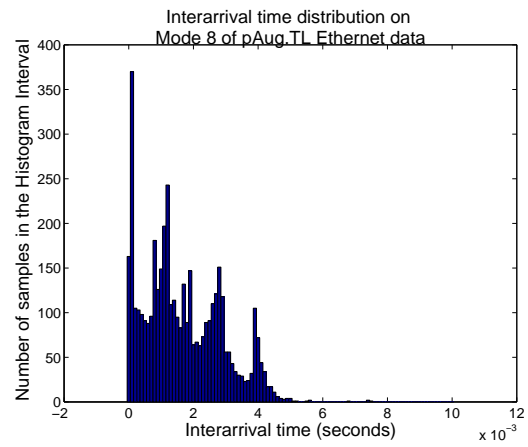
Mode 8

Figure B.3: Interarrival time distributions-Modes 4-8 of pAug.TL trace
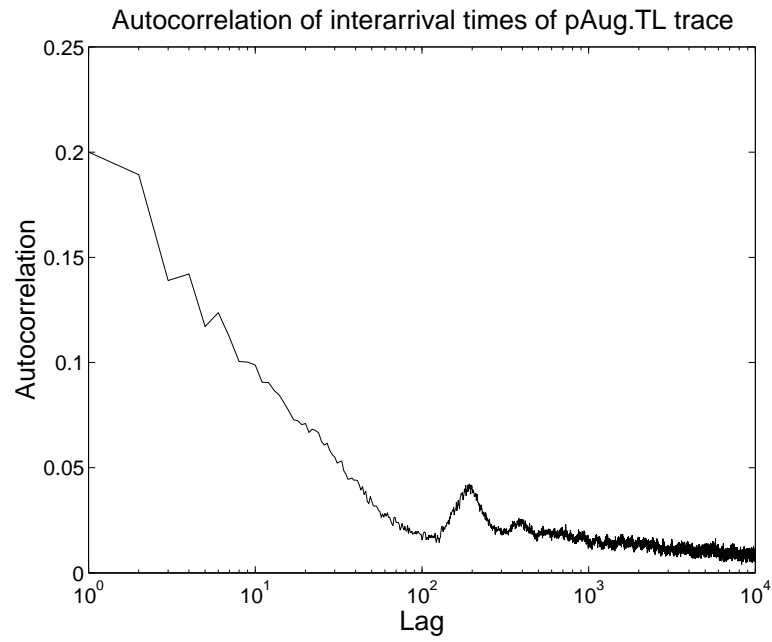
Figure B.4: Lag-k autocorrelation: pAug.TL
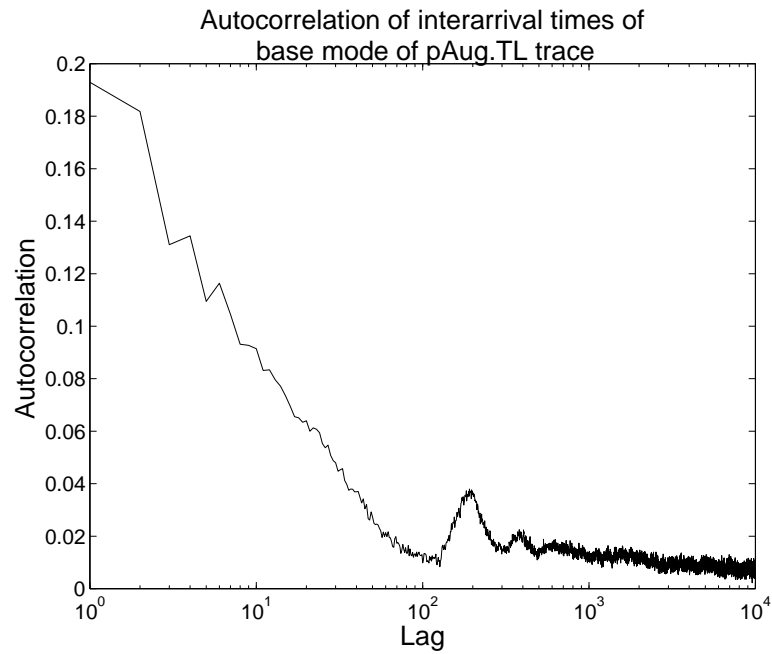
Figure B.5: Lag-k autocorrelation: Base mode
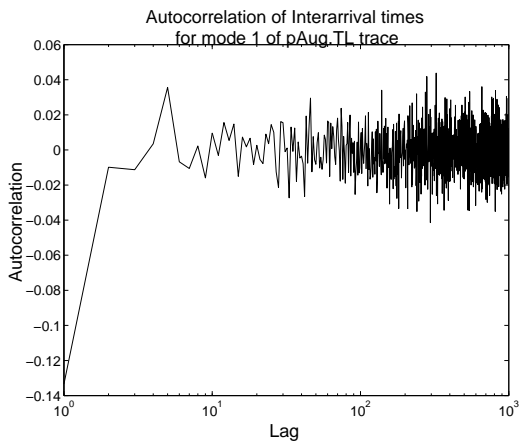
Mode 1
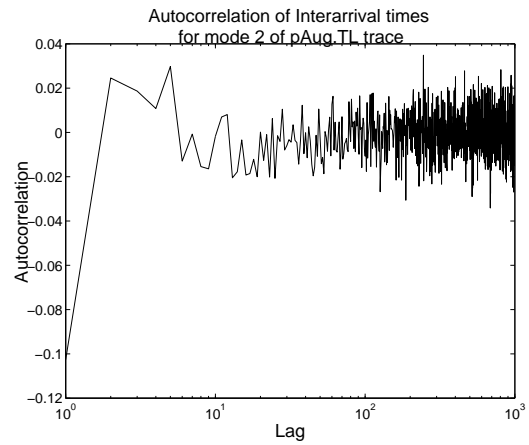
Mode 2
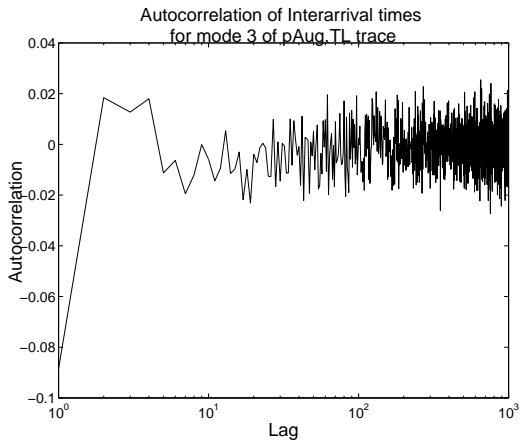
Mode 3
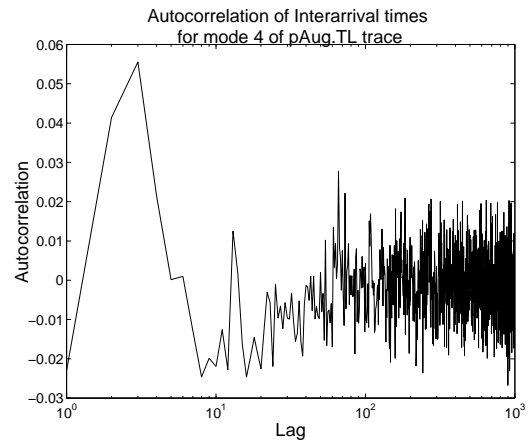
Mode 4

Figure B.6: Lag-k autocorrelations -Modes 1-4 of pAug.TL trace

Mode 5

Mode 6

Mode 7

Mode 8
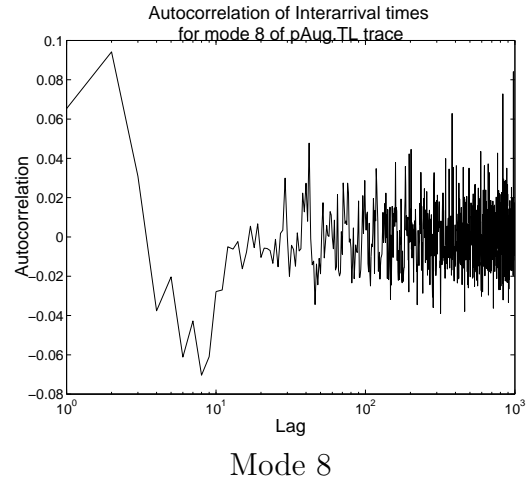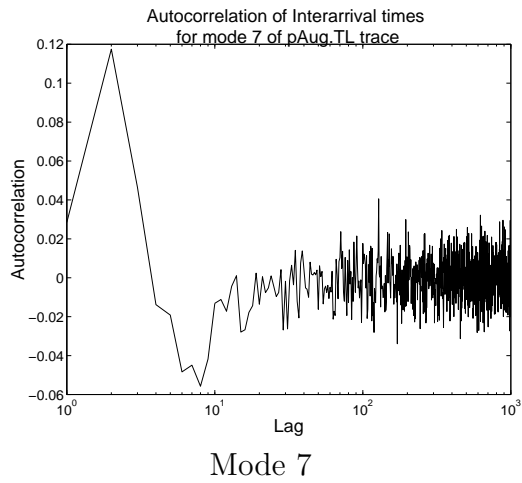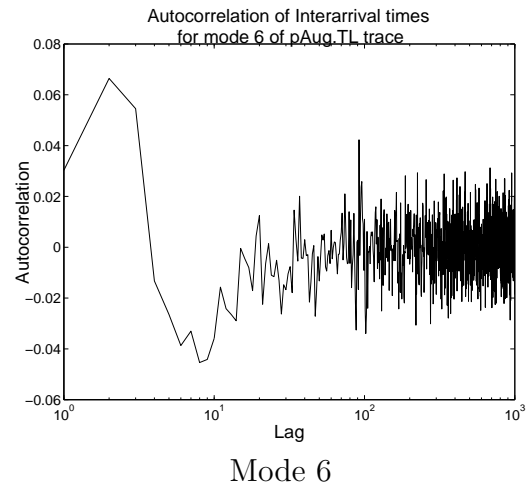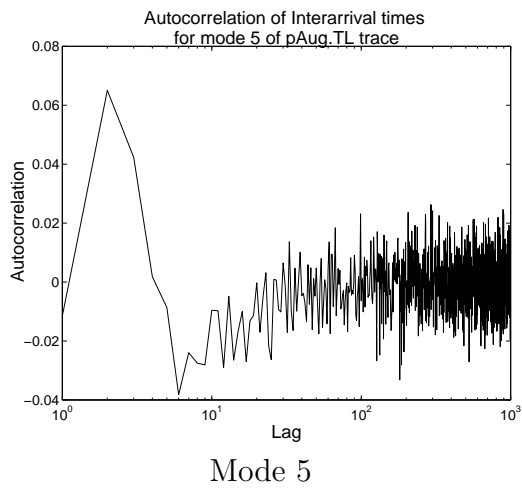
Figure B.7: Lag-k autocorrelations-Modes 4-8 of pAug.TL trace
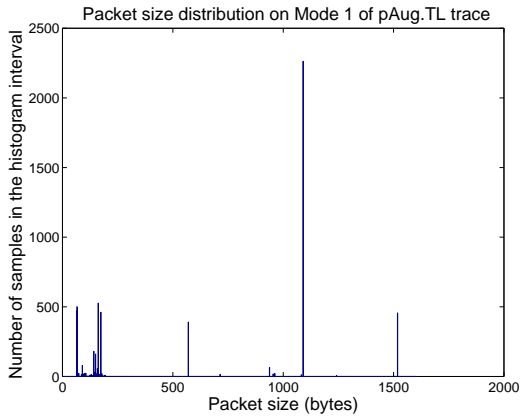
Mode 1
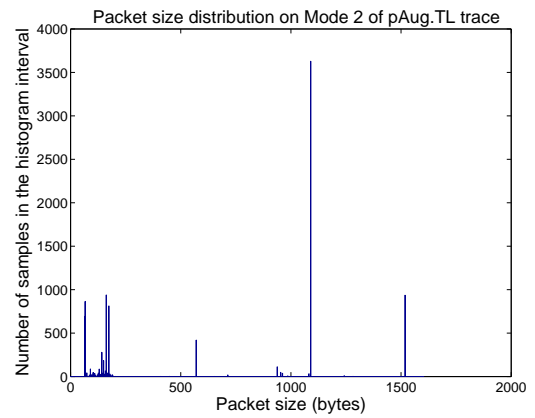
Mode 2
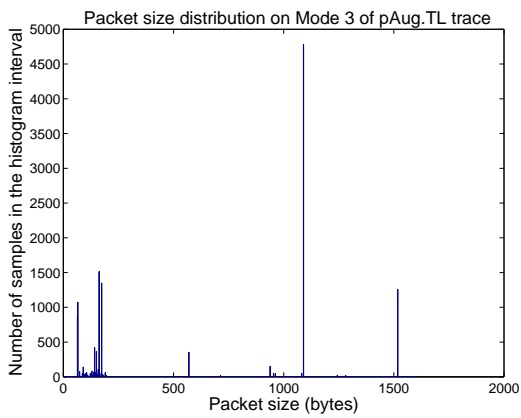
Mode 3
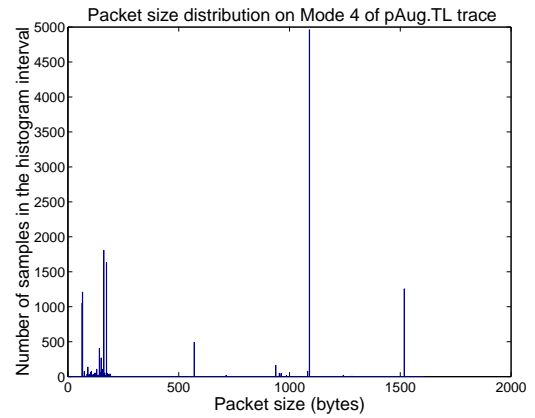
Mode 4

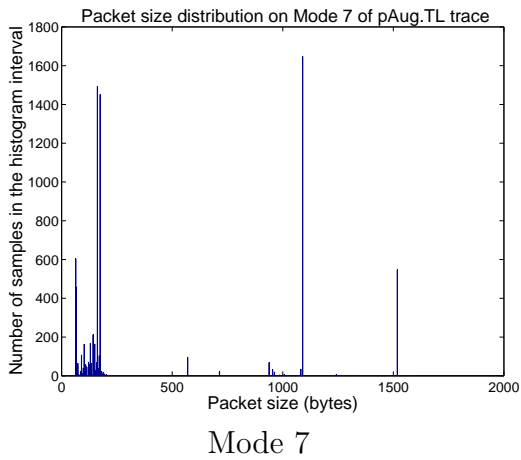Figure B.8: Packet size distribution-Modes 1-4 of pAug.TL trace
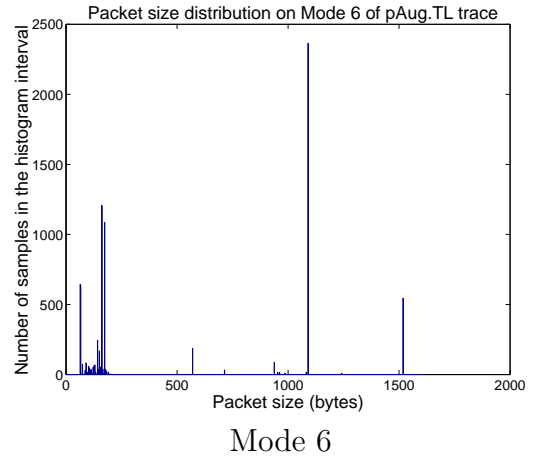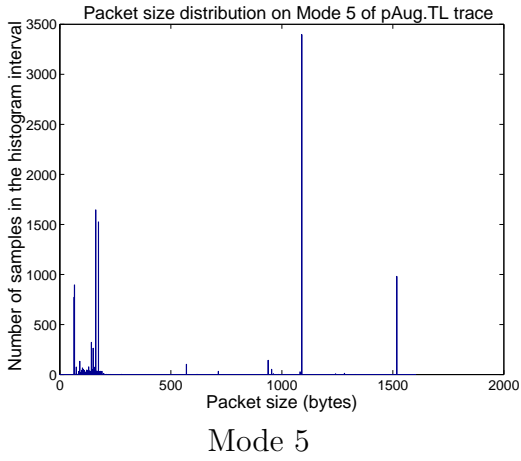
Mode 5

Mode 6

Mode 7

Mode 8
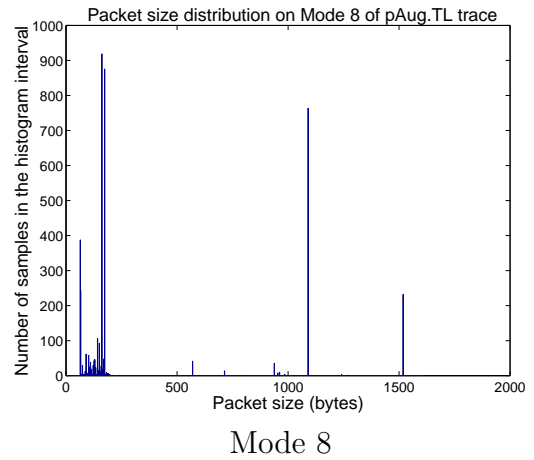
Figure B.9: Packet size distribution-Modes 4-8 of pAug.TL trace