The Effects of Multiple Time-Scale Burstiness and Long-Range Dependence in VBR Video Traffic on Traffic Control in Multimedia Networks
The Effects of Multiple Time-Scale Burstiness and Long-Range Dependence in VBR Video Traffic on Traffic Control in Multimedia Networks

Advisor : Professor Jae-Kyoon Kim

by

Heejune Ahn
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology

A thesis submitted to the faculty of the Korea Advanced Institute of Science and Technology in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical Engineering

Taejon, Korea
1999. 12. 5.
Approved by

[Signature]
Professor Jae-kyoon Kim
Major Advisor
가변비트율 영상트래픽의 다시간 버스트 특성과 장구간 상관성이 멀티미디어 네트워크의 트래픽제어에 미치는 영향

안 희 준

위 논문은 한국과학기술원 박사학위논문으로 학위논문심사위원회에서 심사 통과하였음.

1999 년 12 월 5 일

심사위원장 김 재 균
심사위원 성 단 근
심사위원 조 동 호
심사위원 채 경철
심사위원 정 송
요 약 문 (Abstract in Korean)

기변비트율 영상트레픽의 다시간 버스트 특성과 장구간 상관성이 멀티미디어 네트워크의 트레픽제어에 미치는 영향

영상트레픽은 ATM망, 인터넷, 고속무선망과 같은 멀티미디어 네트워크에서 주요한 부하가 될 것이다. 그러나 영상 트레픽의 복잡한 버스트특성과 깨달고운 전송품질 요구는 효과적인 트래픽 제어를 어렵게 만들고 있다. 최근 다시간 버스트특성과 장구간 상관성으로 정의되는 영상트레픽의 특성이 네트워크 성능에 크게 영향을 주는 것으로 알려져서 연구의 집중이 되고있다.

본 논문은 영상트래픽의 트래픽 제어와 네트워크 설계의 관점에서, 버스트의 시간 준위와 다중와 시스템의 상관관계를 연구하였다. 주요한 연구결과는 장구간 상관성 트래픽 모델, Shifting-Level 과정과 이를 입력으로하는 대기행렬 방식을 제안한 것과, 트래픽제어 결정적인 영향을 주는 cutoff 준위(time-scale)와 dominant 준위의 개념과 호출락제어와 사용자변수제어에서의 응용을 제안한 것이다.

이 논문은 우선, 영상트레픽의 다시간 버스트특성과 장구간상관성의 특성을 트래픽 제어의 관점에서 재평가하고, 트래픽 제어의 국제표준 및 최근 연구들의 상관성을 살펴보는 것에서부터 시작한다. 첫번째 결과로, 장구간 상관성모델인 LRD shifting-level (SL) 확률과정을 제안하고 영상트레픽과의 변수매칭밀고리즘을 제안하였다. 이 SL과정은 네트워크에 영향을 주는 중요한 특성들, 입력분포와, 단구간, 장구간 상관성을 매칭시키면서도, 수학적 분석(양자화 방식으로 명명)이 가능하다는 장점을 갖고있다. 구체적인 장점으로는 (1) 기존의 모델들과 달리, 전구간에 걸쳐 (2) 주어진 오차한도에서 대기행사 과정을 제공하고, (3) 입력분포와 자기상관보다 독립적으로 매칭시킬 수 있다. 특히, 대부분의 영상트레픽이 단구간에서는 지하하향으로, 장구간에서는 쌍곡선함수로 근사화됨을 보였고, 이를 이용하여 각 단구간, 장구간의 상관도가 대기시스템의 성능에 미치는 영향을 분석하였다.

둘째로, 다중화시스템의 성능을 예측하는데 있어서 최대 입력상관도가 되는 cutoff interval의 개념을 도입하였다. 이는 정적, 동적 호출락제어에 필요한 입력트래픽의 상관도의 범위를 정의한다. 또한, 입력곡선을 이용-불록한 곡선으로 근사하는 방안을 제안하여, cutoff time-scale에 관련된 계산을 간단히 하였다. 특히, MPEG/JPEG으로 부호화된 실제 데이터를 사용한 많은 실험을 통하여 LRD의 중요성이 critical time-scale의 위치에 따라 변동하는 것을 보였다. 또한 cutoff time-scale 개념을 SL 모델과 연결하여 트래픽 제어에서의 LRD의 영향을 간단명료하게 설명하였다.

마지막으로, 입력상관도에서 대역할당에 결정적인 영향을 주는 시간 준위인 dominant time-scale의 개념을 도입하고, 이를 바탕으로 표준 UPC변수를 효과적으로 설정하는 방식을 제안하였다. 제안된 방식은 단지 두개의 리커버킷을 사용하는 표준UPC와 많은 수의 리커버킷이 필요한 영상트레픽의 다시간 버스트 특성 사이의 상반된 요구를, dominant time-scale에 충돌을 두고 근사함으로써 수용한다. 여러 가지 MPEG/JPEG 트래스를 사용한 결과들은 critical time의 중요성과 제안된 변수설정방식의 성능의 우수함을 보였다.
Video traffic will be a major load for future integrated services multimedia networks such as ATM networks, the Internet, and high-speed wireless networks. However, the bursty nature and strict Quality of Service (QoS) requirements of video traffic have been obstacles to efficient video transmission. Particularly, two properties of multiple time-scale burstiness and Long-Range Dependence (LRD) in VBR video traffic have been recently reported and are considered to significantly affect network performance.

In this dissertation, we address these two issues from the perspective of traffic control and network design for video transmission by investigating the relationship between the time scale in the correlation of video traffic and the queue buildup. Our main contributions are the proposal of a flexible and mathematically tractable LRD video traffic model, and introduction of two key time scales, cutoff and dominant time-scale which play a key role in Call Admission Control (CAC) and Usage Parameter Control (UPC).

We begin this dissertation by investigating the statistics and origin of the multiple time scale burstiness and LRD of VBR video traffic and relating with the current research activities and the status of traffic control standards.

As the first result, we present a LRD video traffic model based on the shifting-level (SL) process with an accurate parameter matching algorithm for video traffic. The SL process has its strength that it captures all key statistics of an empirical video trace, i.e., short- and long-term correlations and rate-distribution, while still retaining mathematical tractability. We devise a queueing analysis method of SL loaded system, named quantization reduction method. This method provides queueing results over all range of queue size, not just an asymptotic solution. Especially, we found that for most available traces its ACF can be accurately modeled by a compound correlation (SLCC): an exponential function in short range and a hyperbolic function in long range. Through extensive numerical experiments, we identify the effects of SRD and LRD in VBR video traffic on queueing performance.
Second, we introduce the concept of cutoff time-scale, the upper bound of the time scale which affects queue buildup. We show that the cutoff time-scale is an increasing function of delay bound and utilization. This theory provides a fundamental guideline for traffic characterization of video traffic for CAC. Extensive experiment with MPEG/JPEG-compressed video shows that the significance of LRD depends on the location the dominant time-scale, which explains the origin of contradictory arguments in the literature. We also present an algorithm that approximates the arrival bound curve with a concave function. This significantly simplifies the calculation in the estimation of the cutoff time-scale and delay bound with little estimation loss.

Finally, we introduce the concept of dominant time-scale and apply it into the standard-compatible UPC parameters selection. While the standard UPC enforcement is based on the dual leaky bucket, i.e., two leaky buckets, the multiple time-scale burstiness requires a large number of leaky buckets for specifying input traffic. The proposed algorithm fills the gap between the traffic specification standards and real characteristics of traffic, by approximating the traffic arrival pattern accurately at the most important time scale, i.e., the dominant time-scale, for call admission decision. Simulation results with MPEG/JPEG compressed video trace show the efficiency of the proposed algorithm in efficient resource utilization.
Contents

1 Introduction ........................................ 1
   1.1 Video Transmission over Integrated Services Networks .... 1
   1.2 Motivation ...................................... 3
   1.3 Organization of the Dissertation ................... 4

2 Understanding of Video Traffic Characteristics ........... 7
   2.1 Introduction .................................... 7
   2.2 Long-Range Dependence of Video Traffic ............... 8
   2.3 Multiple Time-Scale Burstiness in VBR Video Traffic .... 10
      2.3.1 Statistics in Scene Level ..................... 12
      2.3.2 Statistics in Frame-level ..................... 16
   2.4 Various Approaches for Video Transmission .......... 22
      2.4.1 U-VBR (Unconstrained-VBR) .................. 23
      2.4.2 CBR (Constant Bit Rate) .................... 24
      2.4.3 C-VBR (Constrained-VBR) ................. 24
      2.4.4 A-VBR (Adaptive-VBR) .................... 25

3 The Effects of SRD and LRD on Queueing Behavior: A Video Traffic Model based on the Shifting-Level Process and its Queueing Analysis .................. 28
   3.1 Introduction .................................... 28
   3.2 The Shifting-Level Process ....................... 31
      3.2.1 The Shifting-Level Process and its Properties .... 31
      3.2.2 The SL process as a Video Traffic Model .......... 33
      3.2.3 The SLCC and the C-DAR(1) Model ............. 34
   3.3 Parameter Matching of the SL Process ............... 36
      3.3.1 The Proposed Parameter Matching Algorithm .... 36
      3.3.2 Numerical Results for Empirical Traces ....... 39
   3.4 An Efficient Analysis of the SL/D/1/K Queueing System .... 42
      3.4.1 The Quantization Reduction Method ............. 42
      3.4.2 The Efficiency of the Quantization Reduction Method ... 50
3.5 Queueing Performance Results ........................................ 51
3.6 Chapter Summary ......................................................... 57

4 The Cutoff Time-Scale: the Revelant Time Scales of Video Traffic
for Queueing Behavior ....................................................... 61
4.1 Introduction ............................................................... 61
4.2 The PLAEF: A Deterministic Traffic Characterization .......... 65
4.3 Range of Correlation Relevant to Delay Bound Estimation .... 68
4.4 Application of Cutoff Time-scale to VBR Video Traffic ....... 71
4.5 Approximation of PLAEF by a Concave Function ............... 75
4.6 Chapter Summary .......................................................... 82

5 Dominant Time-Scale and its Application to Standard-Compatible
UPC Selection ............................................................... 85
5.1 Introduction ............................................................... 85
5.2 Dominant Time-Scale in Deterministic Modeling ................. 86
5.3 Leaky Buckets Parameter Selection Algorithm ................... 87
5.4 Numerical Results ....................................................... 91
5.5 Chapter Summary ......................................................... 93

6 Conclusions ................................................................. 96
6.1 Contributions of the Dissertation .................................... 96
6.2 Suggestion for Future Work ............................................ 98
List of Figures

2.1 ACF of empirical trace and short-range and long-range matching ... 11
2.2 Variance time function and its asymptotical slope 11
2.3 Scene characteristics of Star Wars trace 12
2.4 Probability density function of Pareto and the scene length of Star Wars trace 15
2.5 Complementary cumulative distributions of Pareto and the scene length of Star Wars trace 15
2.6 3 types coded Star Wars trace: JPEG, H.261, and MPEG-1 17
2.7 MPEG GOP structure 19
2.8 The bit-rate distribution of I, P, and B picture 20
2.9 Complementary Distribution of JPEG-coded Star Wars trace 20
2.10 ACF for MPEG video traffic 21
2.11 Various types of video transmission in the recent literature: U-VBR, CBR, C-VBR, A-VBR 23

3.1 A sample path for the SLCC with the same ACF and the marginal distribution as the intra-coded Star Wars trace 34
3.2 The ACFs for original, exponential, hyperbolic and SLCC (JPEG) 40
3.3 The ACFs for original, exponential, hyperbolic and SLCC (GOP smoothed MPEG) 40
3.4 The ACFs for original, exponential, hyperbolic and SLCC (VIC Trace A, VIC trace B) 41
3.5 Distribution matching (JPEG Star Wars) 45
3.6 The quantization functions: \( \mathcal{T}^a \), \( \mathcal{T}^u \), and \( \mathcal{T}^l \) 46
3.7 Sample paths for the \( L(t) \), \( L^a(t) \), \( L^u(t) \), and \( L^l(t) \) 47
3.8 Queue occupancies for the empirical, SLCC, and C-DAR(1) inputs 55
3.9 Cell loss probabilities for the empirical, SLCC, and C-DAR(1) inputs 55
3.10 Queue tail: 64 multiplexed GOP smoothed MPEG (Top), VIC trace A (Middle), and B (Bottom), utilization of 0.8 58

4.1 The PLEAF, mean arrival, peak arrival functions for the MPEG trace 69
4.2 The PLAEP, mean arrival, peak arrival functions for the JPEG trace 69
4.3 The illustration of the cutoff time-scale, comparing heavy load and
light load condition ($C_H < C_L$) ........................................... 72
4.4 The cutoff time-scale variation for the MPEG trace against the max-
imum delay bound .......................................................... 74
4.5 The cutoff time-scale variation for the JPEG trace against the maxi-
mum delay bound .......................................................... 74
4.6 Cutoff time-scale versus the mean load of the multiplexer, for various
input traffic mixes. ......................................................... 75
4.7 The Approximation algorithm of an original PLAEF to a concave one 76
4.8 The concave PLAEF and original PLAEF for the MPEG trace ........ 78
4.9 The concave PLAEF and original PLAEF for the JPEG trace ........ 78
4.10 The changes of cutoff time-scales due to the concave approximation
for the MPEG trace against the maximum delay bound ............... 80
4.11 The changes of cutoff time-scales due to the concave approximation
for the JPEG trace against the maximum delay bound ............... 80

5.1 The definition of the dominant time-scale $\tau_d$ in the deterministic traf-
fic specification ............................................................ 88
5.2 The dominant time-scale $\tau_d$ varying the delay bound ............... 89
5.3 Examples for the proposed dual leaky bucket selection algorithm .... 92
5.4 The Wrege's algorithm for multiple leaky bucket approximation .... 93
5.5 Illustration for limit of characterizing input to dual leaky buckets
without the dominant time-scale concept. ................................ 94
5.6 Bandwidth utilization gain of the proposed algorithm ............... 95
List of Tables

2.1 Multiple time-scale burstiness: origin and time-scales
2.2 Statistics for MPEG-coded Star Wars
2.3 Trade-offs of U-, C-, A-VBRs, and CBR video transmission

3.1 Statistics for the traces used in our experiment (cells/frame or cells/GOP)
3.2 Normalized deviations for negative cumulative probabilities: Real trace, Negative Binomial and Gaussian approximation
3.3 More Numerical results on normalized deviations
3.4 Queue occupancies for \( L^u(t) \), \( L^a(t) \), and \( L^l(t) \) (\( U = 0.8 \))
3.5 Queue occupancies for \( L^u(t) \), \( L^a(t) \), and \( L^l(t) \) (\( U = 0.6 \))
3.6 Queue occupancies for \( L^u(t) \), \( L^a(t) \), and \( L^l(t) \) (\( U = 0.4 \))
3.7 Cell loss probabilities for \( L^u(t) \), \( L^a(t) \), and \( L^l(t) \) (\( U=0.8 \))
3.8 Cell loss probabilities for \( L^u(t) \), \( L^a(t) \), and \( L^l(t) \) (\( U = 0.6 \))
3.9 Cell loss probabilities for \( L^u(t) \), \( L^a(t) \), and \( L^l(t) \) (\( U=0.4 \))
3.10 Cell loss probabilities for superposed sources (5, 10) at \( U = 0.8 \)

4.1 Summary of the VBR traces used in the Study
4.2 The Numbers of IAPs \( r \) required to cover upto the cutoff time-scale: \( r \) for the concave PLAEFs ( \( r \) for the original PLAEFs)
4.3 Simulation results for various video traces: cutoff time-scales, \( t_r \), \( \overline{t_r} \) (frames), and link utilization
4.3 Simulation results for various video traces: cutoff time-scales, \( t_r \), \( \overline{t_r} \) (frames), and link utilization (continued)
Chapter 1

Introduction

1.1 Video Transmission over Integrated Services Networks

Multimedia communications will be the first feature of communications in the new millennium, at least the new century. The remarkable growth in development and standardization activities in signal processing and communication technology in the past decade have made multimedia communications possible. There is also rapidly increasing demands of multimedia service from the Internet and wireless networks. Multimedia communications require the capability of networks of carrying various information styles, such as voice, text, video, world-wide Web (WWW), not-yet-come, and their mixture. Due to the diversity in traffic characteristics and Quality of Service (QoS) requirements, multimedia communication networks rely on the packet switching mechanism, where bandwidth is not dedicated to a session and shared by multiple sessions.

Among those multimedia traffics, video traffic provides a serious challenge to the traffic control in multimedia integrated service networks, because of its large bandwidth demand, strict QoS requirement, and complicated burstiness property. Thus, the study on video traffic control has been a major issue in the past decade. For
state-of-the-art of traffic control in the literature, see [95, 61]. Also, the international standardization bodies have been working on the traffic control issues. ITU-T and the ATM Forum [52, 10] for ATM based broadband network To meet the QoS requirements of connections of a variety of imagined types, a number of transfer capabilities have been specified (for more details, see [95] page 55-59). Similarly the Internet Engineering Task Force (IETF) is in the process of defining new service class with guaranteed performance, and integrated service architecture to augment the current ubiquitous “best effort” IP datagram service class: IntServ, and DiffServ [119].

In general, traffic control techniques are classified as either preventive or reactive control. Reactive control has been precluded from traffic control mechanisms in high speed networks since the early stage of study of broadband services, because its stability could not be guaranteed due to the large product of bandwidth and round-trip time (see Chapter 7 in [107]). However, recently it is verified through experiments and theory that the feedback mechanism can improve service quality significantly [56, 65, 39]. Nevertheless, due to the instability question in reactive control, great care should be taken in designing a control algorithm. Further even though reactive control could improve its service a lot, it is very certain that QoS guarantee mainly should be based on preventive control mechanisms.

For these reasons, in this dissertation, we will focus on the preventive control issues. In preventive traffic control, the major traffic control mechanisms are call admission control (CAC) and usage parameter control (UPC). At call-setup, the traffic source first declares its expected usage and required QoS to the network. Then the network calculates the network performance with a new call, and accept the call if it can provide the required QoS without violating the other connections
already in service. Otherwise, the call is rejected or renegotiated (CAC). After call setup, the network polices the usage conformance of the source (UPC).

1.2 Motivation

All preventive control mechanisms, notably, CAC, UPC, and network dimensioning, inherently depend on performance prediction, and the success of this prediction is absolutely governed by the traffic characteristics. Thus, the complex traffic characteristics in video traffic have been a strong challenge against traffic control design. Particularly, two characteristics in a compressed video source have been recently introduced (since early '90s): multiple time scale burstiness and [37, 63, 110] and long-range dependence (LRD) property [38, 16]. Further, these characteristics are considered to play a key role in traffic control. First, LRD means the strong persistence in time-correlation of video traffic. Differently from traditional video traffic models based on Markov process, LRD models results in hyperbolic queue tails in the asymptotic queue size region. These results have aroused the anxious on the failure of Markovian model, i.e., and the possibility of overestimation in network performance [32, 83]. Second, VBR video traffic shows burstiness not only over a period of milliseconds to seconds, corresponding to variations within a scene, but also over a period of tens of seconds to minutes, corresponding to scenes with differing information contents.

Since VBR coded video traffics are quite complex, it is extremely difficult to devise a simple and accurate traffic model. Further, every detail of traffic is not important in the network performance and the design of traffic control. Therefore, it is natural to consider the following problem:

What is the important characteristics in video traffic for traffic control?
In this dissertation, we deal very this issue with theory and extensive experiments. In particular, we will focus on which time-scale affects the performance of multiplexer, i.e., queue buildup, and which time scale determines its UPC selection for QoS guarantee., and how much and when the LRD in video traffic is significant in multiplexer performance.

1.3 Organization of the Dissertation

In this dissertation, we address multiple time-scale burstiness and LRD in video traffic. Our approach is quite unique in that we use top-down approach, i.e., from the perspective of traffic control and network design to video traffic charcterization. Instead of trying to catch all characteristics in video traffic, we investigate the relationship between the time scale in the correlation of video traffic and the queue buildup. Our main contributions are the proposal of a flexible and mathematically tractable LRD video traffic model, and introduction of two key time scales, cutoff and dominant time-scale which play a key role in Call Admission Control (CAC) and Usage Parameter Control (UPC). Whereas our approach is top-down manner, the organization of the dissertation is step by step.

In Chapter 2, we begin this dissertation by investigating the statistics and origin of the multiple time scale burstiness and LRD of VBR video traffic and relating with the current research activities and the status of traffic control standards.

In Chapter 3, we present a stochastic video traffic model based on the shifting-level (SL) process. In spite of long history of video traffic modeling, our results has from other modeling approaches. The SL process captures all key statistics of an empirical video trace, i.e., short- and long-term correlations and rate-distribution. Further, the SL process still retains good mathematical tractability. We devised pa-
rameter matching algorithm where the rate distribution and ACF is independently matched, and thus not only simple and accurate, but also providing the convenience in understanding the effects of each statistic. We devise a queueing analysis method of SL loaded system, named quantization reduction method. This method provides queueing results over all range of queue size, not just an asymptotic solution. Especially, we found that for most available traces its ACF can be accurately modeled by a compound correlation (SLCC): an exponential function in short range and a hyperbolic function in long range. Through extensive numerical experiments, we identify the effects of SRD and LRD in VBR video traffic on queueing performance. The with an accurate parameter matching algorithm for video traffic.

With the understanding from the previous chapters, we tackle CAC and UPC issues in the following chapters.

In Chapter 4, we introduce the concept of cutoff time-scale, which define the upper bound of the time scale which affects queue buildup. We show that the cutoff time-scale is an increasing function of delay bound and utilization. This theory provides a fundamental guideline for traffic characterization of video traffic for CAC. We provides extensive experiment results that shows that the significance of LRD depends on the location the dominant time-scale, which explains the origin of contradictory arguments in the literature. We also present an algorithm that approximates the arrival bound curve with a concave function. This significantly simplifies the calculation in the estimation of the cutoff time-scale and delay bound with little estimation loss.

In Chapter 5, we introduce the concept of dominant time-scale and apply it into the standard-compatible UPC parameters selection. While the standard UPC enforcement is based on the dual leaky bucket, i.e., two leaky buckets, the multiple
time-scale burstiness requires a large number of leaky buckets for specifying input traffic. The proposed algorithm fills the gap between the traffic specification standards and real characteristics of traffic, by approximating the traffic arrival pattern accurately at the most important time scale, i.e., the dominant time-scale, for call admission decision. We also provide simulation results that show the efficiency of the proposed algorithm in efficient resource utilization.

Finally, in Chapter 6 we summarize the contributions of dissertation and indicate future research directions.
Chapter 2

Understanding of Video Traffic Characteristics

2.1 Introduction

From the early period of packet video, i.e., VBR video, the correlation structure of video traffic is intensively studied. Recently, data sets of long full-motion entertainment video have become available to the video traffic researchers. Based on in-depth statistical analysis of these data sets, recent research results can be summarized as follows:

- The histogram of video trace is important in queueing performance [48, 108].

- Video traffic shows burstiness in multiple time-scale [3, 110, 35, 111].

- Long-range dependence (LRD) is an inherent features of VBR video traffic [16, 38, 82].

- Video coding affects the characteristics of video traffic [62, 100, 101].

This chapter provides the background for understanding the following chapters. First, we introduce the notion of LRD (long-range dependence) in comparison with
SRD (short-range dependence), a brief overview on the issues and approaches for transmitting video traffic over integrated services networks.

In first part, we introduce long-range dependence and multiple time-scale burstiness in VBR video traffic, and then consider transmission mechanisms for VBR video.

The variation of video traffic depends on both visual information and coding algorithms. The slow-varying components of video traffic is due to inherent characteristics of visual information, the scene activities and the resolution of visual components.

The multiple time-scale burstiness and long-range dependence introduced a new challenge in video traffic control. First, LRD is defined by infinite summation of autocorrelation function and its evidence in VBR video traffic has been verified through real traces. It has been reported that LRD could cause network performances totally different from those of the present Markov Model [32, 82, 38]

Second, the motivation of multiple time-scale burstiness comes from understanding the visual information of entertainment videos. A video sequence contains scenes, frames, and slices and each information unit is different in (1) origin, (2) the statistics, and (3) time-scale. Therefore, it is natural to think that the effect on the network performance is different and the control algorithm efficient for each uses different mechanism.

2.2 Long-Range Dependence of Video Traffic

Intuitively, LRD, also known as "persistence" or the "Hurst effect," is the phenomenon of empirical records being significantly correlated to records far away in time.
The correlation structure of video traffic is well-described by auto-correlation function (ACF) \( \rho(k)(k = 0, 1, 2, \cdots) \). A process with summable ACF (\( \sum_k \rho(k) < \infty \)) is called a SRD process, and, in contrast, a process with infinite sum of ACF (\( \sum_k \rho(k) = \infty \)) is called as a LRD process. Let \( X(t), t = 0, 1, 2, \cdots \), be a covariance stationary stochastic process with mean \( m \), variance \( \sigma^2 \) and auto-correlation function (ACF) \( \rho(k) \). \( X(t) \) is called a LRD process, if

\[
\rho(k) \sim k^{-\beta} L(k), \quad \text{as } k \to \infty, \tag{2.1}
\]

where \( 0 < \beta < 1 \), and \( L \) is a slowly varying function, that is, \( \lim_{t \to \infty} L(tx)/L(t) = 1 \), for all \( x > 0 \). Constants and logarithms are examples of slowly varying function. For each \( m = 1, 2, \cdots \), let \( X^{(m)}(k), k = 1, 2, \cdots \) be the new covariance stationary process defined by

\[
X^{(m)}(k) = (X(km - m + 1) + \cdots + X(km))/m. \tag{2.2}
\]

Let \( \text{Var}(X^{(m)}) \) denote the variance of \( X^{(m)} \). The process \( X(t) \) is called strictly second-order self-similar with self-similarity parameter \( H = 1 - \beta/2 \) if for all \( m \geq 1, \sigma_m^2 = \sigma^2 m^{-\beta} \). If

\[
\sigma^{(m)} \to \text{const} \cdot m^{-\beta} \quad \text{as } m \to \infty, \tag{2.3}
\]

the process \( X_t \) is called (asymptotically) second-order self-similar with self-similarity parameter \( H = 1 - \beta/2 (\frac{1}{2} < H < 1) \).

A few estimation methods for LRD and its Hurst parameter have been devised: the variance-time plot, R/S analysis, and periodogram (i.e., power spectral density). Variance-time plot is a trace of \((\log_{10} m, \log_{10} \text{var}(X^{(m)}))\).

VBR video traffic has both SRD and LRD. Since most traffic models has only one of SRD and LRD, the effect of SRD and LRD on traffic control is separately studied.
Table 2.1: Multiple time-scale burstiness: origin and time-scales

<table>
<thead>
<tr>
<th>scene</th>
<th>frame &amp; slices</th>
<th>cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>origin</td>
<td>visual content</td>
<td>coding schemes</td>
</tr>
<tr>
<td>time-scale</td>
<td>a few sec to min</td>
<td>several msec</td>
</tr>
</tbody>
</table>

In Chapter 3, we will investigate the effects of the SRD and LRD in a seamless way. Figure 2.1 shows that, in $n > 100$, the ACF of JPEG-coded *Star Wars* trace can be modeled by the hyperbolic function. Figure 2.2 shows the variance-time curve for *Star Wars* trace. According to the definitions, VBR video traffics is not strictly but asymptotically self-similar process with $\beta = 0.38$. Furthermore, the region less than $m = 100$ shows very strong correlation.

2.3 Multiple Time-Scale Burstiness in VBR Video Traffic

The motivation of *multiple time-scale burstiness* comes from understanding the visual information of entertainment videos. A video sequence contains *scenes, frames, and slices* and each information unit is different in (1) origin, (2) the statistics, and (3) time-scale. Therefore, it is natural to think that the effect on the network performance is different and the control algorithm efficient for each uses different mechanism.

Figure 2.3 shows the multiple time-scale burstiness, emphasizing the scenic process.
Figure 2.1: ACF of empirical trace and short-range and long-range matching

Figure 2.2: Variance time function and its asymptotical slope
2.3.1 Statistics in Scene Level

To focus on the inherent characteristics in VBR video traffic, and thus use JPEG coded *Star Wars* video traffic data [38] instead of MPEG traffic which has periodic burstiness. Let's consider a video trace as a successive sequence of scene, and explore the statistics of scenes. The process related with scenes is called as *scenic process*.

We first extract the boundary of scenes, and classify the scenes into a few types of scene based on the average frame size. Since visual scene detection needs real sequence, a methodology similar to that introduced in [67] was used. A scene change is declared when the absolute difference (AD) between successive bit exceeds a certain threshold $J_{\text{min}}$. Let the jump function $J_n$ denote $J_n = 1[|X_n - X_{n-1}| > J_{\text{min}}]$. It is also observed from visual record that the first frame following a scene change usually had a bit rate, uncorrelated to that of the following or preceding, due to the
interlacing effect. The indicator function $SC_n$ is then given by

$$SC_n = 1\{J_n = 1, J_{n-1} \neq 1, \ldots, J_{n-L_{\min}} \neq 1\}. \tag{2.4}$$

In *Star Wars*, the numerical data were obtained by setting $J_{\min} = 5000$ bytes, which is about one fifth of average bit rate, and $L_{\min} = 2$. Figure 2.3 shows the detected scenes from the original *Star Wars* trace.

We then classify the scenes into 4 classes $\lambda_i, i = S, M, L, XL$, corresponding to the input rates $S = \{\bar{\lambda} - 2\sigma < \lambda < \bar{\lambda} - \sigma\}, \ M = \{\bar{\lambda} - \sigma < \lambda < \bar{\lambda} + \sigma\}, \ L = \{\bar{\lambda} + \sigma < \lambda < \bar{\lambda} + 3\sigma\}, \ XL = \{\bar{\lambda} + 3\sigma < \lambda\}$. After classifying the scenes, we merge successive scenes which are the same class.

Based on the scene extraction technique above, we obtain scenic process. The major stochastic process related on scene process is (1) $L_n$, the scene length and (2) $S_n = S, M, L, or, XL$, the scene size process of the n-th scene.

Here we model the Scenic process as a Semi-Markov Process $\{X(t) : t \in \mathbb{R}_+\}$ with state $\{S, M, L, XL\}$ and a semi-Markov kernel $G(\cdot) = (G_{ij}(\cdot))_{i,j=1}^4$. \footnote{Let $t_n, n = 1, 2, \ldots$ denote the n-th state transition time and $X_n$ is the new state at $t_n$. Then, an element of the semi-Markov kernel $G$, $G_{i,j}(t)$ is defined by

$$G_{i,j} = \text{Prob}[t_{n+1} - t_n \leq t | X_{n+1} = j, X_n = i]$$

where $F(t)$ is a distribution function and $G = G(\infty)$ is an irreducible stochastic transition matrix.

First, By counting the relative frequencies of scene transitions, we obtain the...
limiting transition probability matrix
\[
G = \begin{pmatrix}
0 & ps_M & ps_L & ps_{XL} \\
ps_M & 0 & pm_{XL} & pm_{XL} \\
ps_L & pm_L & 0 & pl_{XL} \\
px_{XL} & px_{XL} & px_{XL} & 0
\end{pmatrix} = \begin{pmatrix}
0 & 0.8833 & 0.1011 & 0.0106 \\
0.4637 & 0 & 0.5112 & 0.0251 \\
0.0976 & 0.8927 & 0 & 0.0098 \\
0.0833 & 0.6667 & 0.2500 & 0
\end{pmatrix}
\] (2.6)

and the arrival rates matrix \( \Lambda = \text{diag}\{\lambda_S, \lambda_M \lambda_L, \lambda_{XL}\} \) bytes/frame is given by
\[
(1.6160e + 004, 2.7763e + 004, 3.9367e + 004, 5.0970e + 004).
\] (2.7)

Here the \( \lambda_s \)s are obtained by averaging the frame sizes of each scene class, respectively.

Now we consider the scene length distribution, \( T \sim F(t) \) with mean \( \mu_t \). In [49], they proposed many candidates for \( F(t) \), including the binomial, Beta, Gamma distributions. In [53], \( F(t) \) is modeled by (classical) Pareto distribution.

\[
F_P(t) = \begin{cases}
1 - \left(\frac{\mu \beta}{(\beta + 1) t}\right)^{\beta + 1} & \text{for } t \geq \frac{\mu \beta}{\beta + 1}, \\
0 & \text{for } t < \frac{\mu \beta}{\beta + 1},
\end{cases}
\] (2.8)

where \( \mu_t \) is mean of the Pareto variable, and \( \beta \) is the tail distribution parameter \( (\beta = 2 - 2H) \). However, here we use a variant of Pareto distribution, called translated Paperto,

\[
F_P(t) = 1 - \left(\frac{\mu \beta}{\mu \beta + t}\right)^{\beta + 1}, \quad 0 < \beta < 1
\] (2.9)

because the discontinuity of classical Pareto often makes some difficulty in calculation.

**Theorem 2.1** [53]: A semi-MMPP process with kernel of Pareto distribution is an asymptotically self-similar process with \( H = 1 - \beta/2 \).

**Proof and Comment:** For Proof, see [53], but briefly speaking, the asymptotic tail of ACF is same as Markov Renewal process with \( F_P(t) \), which has hyperbolic tail ACF. However, we should note that even though the two processes have the
Figure 2.4: Probability density function of Pareto and the scene length of *Star Wars* trace

Figure 2.5: Complementary cumulative distributions of Pareto and the scene length of *Star Wars* trace
same asymptotic tail, but the ACF in the small and middle range will be different from each other.

To obtain the parameters of the translated Pareto distribution, we match the mean $\mu_T$ and tail slope to the empirical distribution ($\beta + 1 = 3 - 2H$). We obtained $\beta = 0.4$ and $\mu_t = 220$ frames. Therefore,

$$F(t) = 1 - \left( \frac{0.4 \cdot 220}{220 \cdot 0.4 + t} \right)^{1.4}.$$  \hspace{1cm} (2.10)

Figures 2.4 and 2.5 show the match of the distribution and the empirical distribution which have the same mean value and tail characteristics.

### 2.3.2 Statistics in Frame-level

We now investigate the frame-level statistics of video traffic. As mentioned above, frame level statistics depend on both the visual information itself and video coding algorithm used for data compression. Figure 2.6 shows parts of video traces coded in 3 different video standards, JPEG, H.261, and MPEG-1.

First, we briefly describes the three popular picture coding standards:

- **JPEG (Joint Picture Expert Group) [55, 112]** is a still picture coding standards. The JPEG algorithm exploits only intra-frame redundancy, thus intra-coding only. Roughly speaking, in the (baseline) JPEG standard, the digitized visual signal, i.e., a picture, is coded through the sequential processing. First, a frame is divided into $8 \times 8$ pixel blocks, which is transformed via the DCT (discrete-time Cosine Transform) transform. Then the DCT coefficients are approximated by an adaptive quantization matrix. Finally, the quantized DCT values are scanned in the zigzag manner and coded using 2-dimensional Huffman coding.
Figure 2.6: 3 types coded Star Wars trace: JPEG, H.261, and MPEG-1
• H.261 standard [43] is a visual coding standards for video-telephony applications. Differently from the JPEG standard, H.261 algorithm exploits inter-frame redundancy also by motion compensation. First, a frame is divided into $16 \times 16$ Macro-blocks, which are composed by $4 \times 4 $ luminance and $2$ chrominance blocks. Each Macroblock is predicted from the previous frame using a motion vector, which denotes displacement vector pointing the most similar macroblock in the previous frame. After motion compensation, the difference signal of the current blocks from the previous blocks are coded in the same manner in JPEG coding, i.e., DCT, quantization, zigzag scan, and Huffman coding.

• MPEG (Moving Picture Experts Group) standards [79] is a generic visual coding algorithm for storage, communication, and broadcasting. The basic algorithm in MPEG coding is also motion-compensation, DCT, quantization, and Huffman coding as in H.261 coding. From the perspective of video transmission, the most salient difference from other standards is its Group of picture (GOP) structure in compression.

Figure 2.7 shows a typical GOP structure for MPEG coding scheme. In the MPEG video coding algorithm, each picture is coded in different coding schemes: Intra (I) coding for random access and error resilience, Predictive (P) coding for normal data compression, and Bi-directionally predictive (B) coding for more efficient data compression. The I-picture generates much larger amount of data than others, typically about ten times that of the B-picture. Furthermore, during one session, the video sequence is usually coded according to a predetermined GOP structure.
Now we present some important statistics at frame level with *Star Wars* trace. Originally, *Star Wars* sequence consisting of 174126 frames was provided by M. Garret [38]. The GOP structure for MPEG trace is "IBBPBBPMBBPBBI...", i.e., \((G,M) = (12,3)\). Here \(G\) represents the number of frames between two consecutive I frames, so called GOP period, and \(M\) represents the number of frames between two P (or I) frames. Table 2.3.2 gives some salient statistical properties for *Star Wars* sequences.

In general histogram and ACF is considered as the most significant components on network performance. It has been reported by several studies that the histogram with fixed quantization scale can be accurately modeled by negative binomial distribution. Figure 2.8 shows the histogram characteristics for I, P, B pictures, comparing with the negative binomial distributions with same means and variances as empirical data. Figure 2.9 illustrate the difference of histogram with JPEG coded *Star Wars*
Figure 2.8: The bit-rate distribution of I, P, and B picture

Figure 2.9: Complementary Distribution of JPEG-coded Star Wars trace
Figure 2.10: ACF for MPEG video traffic

Figure 2.10 shows the empirical ACF of the MPEG-coded video. Comparing the ACF of JPEG coded Star Wars trace in Figure 2.1, the ACF of MPEG trace oscillates heavily and damps slowly; The envelope of ACF decays very slowly in comparison with the oscillation, and the periodic pattern due to the GOP pattern strongly prevails.

Many traffic models for simulation and analysis have been proposed for MPEG coded videos. In [84] and [99], each picture coding type are classified into different states in a Markovian process, and the sojourn time of each state is approximated as exponential distribution. In [81], Nicola has presented a traffic model which captures the periodic burstiness due to the MPEG GOP structure. In [99], we proposed a CAC algorithm considering the picture-dependent arrival pattern in ATM
Table 2.2: Statistics for MPEG-coded *Star Wars*

<table>
<thead>
<tr>
<th></th>
<th>all frames</th>
<th>I</th>
<th>P</th>
<th>B</th>
<th>GOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>15,599</td>
<td>60,416</td>
<td>23,081</td>
<td>7,196</td>
<td>187,185</td>
</tr>
<tr>
<td>peak</td>
<td>185,267</td>
<td>185,267</td>
<td>174,294</td>
<td>64,785</td>
<td>932,710</td>
</tr>
<tr>
<td>peak/mean</td>
<td>5.0</td>
<td>3.06</td>
<td>7.55</td>
<td>9.00</td>
<td>4.98</td>
</tr>
<tr>
<td>std. (Mbps)</td>
<td>18,165</td>
<td>19,805</td>
<td>14,661</td>
<td>4,812</td>
<td>72,468</td>
</tr>
<tr>
<td>Cov. Coef.</td>
<td>1.16</td>
<td>0.32</td>
<td>0.63</td>
<td>0.66</td>
<td>0.39</td>
</tr>
<tr>
<td>min</td>
<td>476</td>
<td>11,754</td>
<td>2,192</td>
<td>476</td>
<td>77,754</td>
</tr>
</tbody>
</table>

multilexer.

2.4 Various Approaches for Video Transmission

The promised advantages of VBR video transmission over CBR transmission have been 1) better video quality for the same average bit rate by avoiding the need to adjust the quantization as in CBR, 2) shorter delay since the encoder buffer size can be reduced without encountering an equivalent delay in the network, and 3) increased link utilization because the bandwidth per a source for VBR video could be lower than for CBR for equivalent quality.

While these potential advantages were heavily emphasized in early studies on VBR video transmission, further research has shown that no design can maximize them simultaneously. Thus, many trade-off approaches can be found in the literatures. In this section we classify and compare trade-offs of various mechanisms based on the classification in [65].

We classify video transmission mechanisms into 4 kind transmission algorithms for video service: unconstrained VBR (U-VBR), constant-bit rate (CBR), constrained VBR (C-VBR), and adaptive VBR (A-VBR) 2.11.
Figure 2.11: Various types of video transmission in the recent literature: U-VBR, CBR, C-VBR, A-VBR

First, we summarize the key features of these mechanism:

- **U-VBR**: unconstrained VBR
- **CBR**: constant bit rate, rate-control against buffer over and underflow.
- **C-VBR**: Constrained VBR, encoder rate-control for UPC.
- **A-VBR**: Adaptive-VBR, network feedback information is used

With the generic system diagram, Figure 2.11, we define 4 transmission mechanisms.

### 2.4.1 U-VBR (Unconstrained-VBR)

The Unconstrained VBR means the VBR transmission In the early stage of VBR research. The video encoder operates independently of buffer and UNI, and network
status, trying to keep consistent video quality during its session. Usually, constant quantization scale is assumed for consistent video quality. Traffic control for QoS guarantees, i.e., maximum cell loss ratio, maximum delay, maximum delay jitter, is most difficult because of no cooperation for congestion control in the end-terminal and the complexity in video traffic burstiness. A totally different but still using CBR channel was presented in [93]. The VBR-coded streams are transmitted with large decoder buffer and constant channel. This approach could be used in broadcasting or near VoD style application with large Disk storage decoder.

2.4.2 CBR (Constant Bit Rate)

In CBR transmission, the encoder adjusts quantization scale in order to meet the bit rate, which is chosen at the connection set-up time. In order to reduce video quality degradation at the high activity or complex frame, the encoder requires an efficient buffer control algorithm and a considerably large buffer, which inevitably introduces extra delay according to the smoothing function. Nevertheless, CBR transmission has popularity because it requires the lowest complexity in traffic control and currently most commercially available networks can support CBR transmission.

2.4.3 C-VBR (Constrained-VBR)

The C-VBR mechanism is based on the notion of traffic contract and preventive control. The C-VBR is often called closed-loop VBR, contrasted with U-VBR called open-loop VBR. The C-VBR encoder controls its rate with defined burstiness constraints, i.e., a contract at the connection set-up time, so that the output is network-friendly, and thus network control is much feasible than U-VBR. The most popular arrival constraints are the single and the dual leaky-bucket control
[10, 52, 109]. Here we briefly describe the leaky-bucket constraint: Leaky bucket is defined by two parameters, leaky rate \( \rho_i \) and bucket size \( \sigma_i \). The bucket counter is incremented at rate \( \rho_i \) upto maximum \( \sigma_i \) and decremented as data are admitted to the network by the corresponding number of bits. We assume that incoming data is discarded, when the counter reaches zero. Thus, the amount of input \( A(s, t) \) arriving over \([s, t]\) is bounded by

\[
A(s, t) < \rho_i(t - s) + \sigma_i
\]  

(2.11)

Reibman and Haskell [90] have studied rate-control under leaky bucket constraint, which approach was followed by many researchers [21, 51, 74, 46]. Pancha and Zarki studied the leaky bucket parameters for MPEG-and JPEG-coded traces [85]. However, C-VBR cannot avoid the picture degradation in high activity scenes in the practical leaky bucket parameters condition [46, 111]. J. W. Roberts, P. Rolin, and M. Hamdi in [46] argue as follows: “We pretend that the full variability of open-loop coding (which is another name of U-VBR) is not necessary to maintain the subjective quality of video sequences containing scenes of different types. Quality from user point of view defends mainly on the visual capacity to capture the information displayed on the screen. In fast moving scenes with complex image structure, the human eye does not have enough time to discover all image details. So, we suggest that the high bit rate generated for such scenes by an open-loop coder is unnecessarily generous.”

2.4.4 A-VBR (Adaptive-VBR)

The A-VBR mechanism is different in the sense that it is based on reactive control and thus utilizes network feedback information.

Most previous studies on video traffic controls are based on preventive congestion
control, which is consistently advocated by emphasizing that reactive control mechanism is not effective in networks with large product of bandwidth and round-trip delay. However, recently feedback control mechanism has revived with successful standardization of ABR and ABT mechanisms [10, 52]. Traffic controls using network status information have been studied in [17, 111, 65, 56, 73, 92]. Lakshman et al. [65, 56] investigate the application of the ABR service on video transmission, and show its possibility.

In A-VBR transmission, the encoder adjust it output bit-rate to the rate available in current network condition. We can further classify the A-VBR into active A-VBR and passive A-VBR: in passive ABR, the encoder controls it bit-rate according to the information from network side, while active A-VBR encoder request predicted bandwidth as well as controls bit-rates according to the allowed bandwidth.

Tse et al. [111] proposed a negotiated CBR (R-CBR) service, where source negotiates and transmits piece-wise linear CBR data. Boyer et al. [17, 42] proposed a reservation mechanism, on which the current ABT transfer capability is. The two mechanisms proposed there is the immediate transfer (IT) and delayed transfer (DT). IT transfer the burst of data with information cell including the cell rate and burst size of the following cells, immediately when it generates. In DT mode, the sender sends a information cell to network and wait for the response of network. Network node can response with positive acknowledgement or negative acknowledgement. From comparison studies of ABT/DT and ABT/IT, it is known that DT is applicable for delay-insensitive service, and IT is for the service that is not so stringent on cell loss.

Finally, we summarize the trade-off of 4 transmission modes in table 2.3. Clearly, CBR makes the task of network management easier because of its predictable traffic
Table 2.3: Trade-offs of U-, C-, A-VBRs, and CBR video transmission

<table>
<thead>
<tr>
<th></th>
<th>CBR</th>
<th>U-VBR</th>
<th>C-VBR</th>
<th>A-VBR</th>
</tr>
</thead>
<tbody>
<tr>
<td>picture</td>
<td>lowest</td>
<td>excellent</td>
<td>moderate</td>
<td>low</td>
</tr>
<tr>
<td>quality</td>
<td>(same avg. BW)</td>
<td>high</td>
<td>poor</td>
<td>moderate</td>
</tr>
<tr>
<td>delay</td>
<td>largest</td>
<td>smallest</td>
<td>small</td>
<td>large</td>
</tr>
<tr>
<td>codec complexity</td>
<td>high</td>
<td>simple</td>
<td>high</td>
<td>simple</td>
</tr>
<tr>
<td>CAC complexity</td>
<td>very simple</td>
<td>very high</td>
<td>moderate</td>
<td>signaling</td>
</tr>
<tr>
<td>mux. gain</td>
<td>very low</td>
<td>high</td>
<td>low</td>
<td>very high</td>
</tr>
</tbody>
</table>

pattern, but it precludes multiplexing gain and thus results in either picture quality degradation or low link utilization. The U-VBR could be ideal in delay and picture quality, but it is considered extremely hard to implement control algorithm to support U-VBR for uncertainty in traffic characteristics. However, the traffic characteristics is still unpredictable, and thus it also experience same problem in CAC as U-VBR.

Consequently, we argue that the C-VBR and A-VBR are the most promising mechanisms for transmitting video traffic over integrated services networks with QoS guarantee. Nevertheless, the A-VBR transmission still has the following problems to solve [1, 34, 106, 111, 121, 122]: 1) how to predict required bandwidth in live video, and 2) how to reduce the signaling cost and 2) how to gracefully adapt the allowed bandwidth on possible failure of bandwidth negotiation.
Chapter 3

The Effects of SRD and LRD on Queueing Behavior: A Video Traffic Model based on the Shifting-Level Process and its Queueing Analysis

3.1 Introduction

This chapter begins the first contribution of the dissertation with a LRD stochastic modeling and queueing analysis, a rather classical approach.

VBR video service is expected to be a major source of future packet-switching integrated service networks. Because the success of traffic control relies essentially on a sound understanding of input traffic, modeling of VBR video traffic has received intense interest. Of input traffic statistics, the histogram (rate-distribution) and the autocorrelation function (ACF), which is equivalent to the power spectrum, are considered of first importance in estimating network performances.

Recently, a number of empirical studies have demonstrated the existence of long-

\footnote{The excerpts from this chapter appear in [6, 7, 9] with different focuses.}
range dependence (LRD) or self-similarity in VBR video traffic [15, 38, 40]. Various processes have been proposed for modeling traffic with LRD and analyzing its effects on network performance [15] and [95] pp.324-348. These include fractional Brownian motion [38, 83], fractional ARIMA processes [50], chaotic maps [31], and semi-Markovian processes [4, 54]. It has been reported that the LRD characteristic has a significant impact on queueing behavior. In particular, in their studies on data traffic, [32] and [83] showed that overall packet loss decreases very slowly with increasing buffer size. In other words, system performance may be overestimated if LRD in input traffic is overlooked.

Nevertheless, conventional models for VBR video traffic are still based on Markovian processes. Markovian models are still being used and developed for performance estimation and traffic control. Furthermore, advocates of Markovian modeling argue that Markovian models show accurate performance estimation in many situations in spite of a lack of LRD characteristics [29, 49]. Therefore, it is natural to consider the following problem [33]: Under what conditions are LRD correlation of VBR video traffic crucial for queueing behavior?

The ambiguity in the argument over the question whether LRD in VBR video traffic has significant effects on the network performance is due to the lack of a traffic model that captures both SRD (in small lags) and LRD (in large lags) characteristics and differentiates the effects of SRD and LRD. Hence we introduce a video traffic model based on the shifting-level (SL) process. The SL process was first studied by Mandelbrot [76] in an economics context and called a renewal reward process. An application to video traffic was done by Grasse et. al. [40]. Roughly speaking, the SL process is a traffic model for a source that changes its rate now and then according to two i.i.d. processes: \( S_t \) for scene size (the arrival rate in a scene) and \( T_t \) for scene
duration. As we will describe, because in the SL process, the histogram and ACF can be matched independently, the effects of each statistic on queueing performance can be investigated separately. With the assumption of a negative binomial distribution on the histogram \([29, 38, 48]\), we focus on the effect of autocorrelation on queueing behavior.

We observe that the ACF of an empirical trace is accurately captured by the shifting-level process with a compound correlation of the exponential and the hyperbolic, which we will refer to as the \textit{shifting-level process with compound correlation} (SLCC). Especially, we present an efficient and accurate parameter matching algorithm for the distributions of \(T_i\) and \(S_i\). The continuous-time analog (C-DAR(1)) of DAR(1) model \([48, 115]\), which is a widely used SRD video traffic model, can be considered a kind of SL process but with only an exponential correlation. Therefore, comparing the queueing performances of the C-DAR(1) model and SLCC with that of a real video trace, we can identify the effects of SRD and LRD correlation in VBR video traffic on queueing performance.

An exact analysis of the SL/D/1/K queueing system is very difficult, because the system size takes a continuous value and the SL process is not Markovian. Hence, we present an approximating method named the \textit{quantization reduction method}, where the system sizes at embedded points, i.e., the rate transition epochs of the SL process, are approximated into a space with a finite number of quantization points. Furthermore, we provide also the upper and lower bounds of the approximation for the system size distribution, and thus we can efficiently use the memory size and computation time by adjusting the quantization step. In addition, for the SL process input, we do not suffer from the computational load because in the SL process the dimension of the one-step transition matrix does not depend on the state size of
the input process. Thus we can accurately match the histogram of an input process without a significant increase in the computational load.

The rest of the chapter is structured as follows. In Section 3.2, we introduce the SL process and its properties for parameter matching. In Section 3.3, we propose a parameter matching method of the SLCC for ACF with exponential and hyperbolic form. In Section 3.4, we offer an analysis method for the SL/D/1/K queueing system. The numerical results on queueing experiments are presented in Section 3.5. Finally we conclude the work in Section 3.6.

3.2 The Shifting-Level Process

3.2.1 The Shifting-Level Process and its Properties

Let $t_0(=0) < t_1 < t_2 < \cdots$ be a renewal process with inter-renewal times $T_n = t_n - t_{n-1}, n = 1, 2, \cdots$, and $S_n, n = 0, 1, 2, \cdots$ be an i.i.d. random process independent of the renewal process $t_n$ with a state space $\{0, 1, 2, \cdots, i, \cdots, M\}$. Furthermore, we assume that the first epoch $t_1$ follows the residual time of $T$. Let $N(t) = \max\{n : t_n < t\}$ denote the index for $t_n$ and $S_n$ at time $t$. Then the shifting-level process $X(t)$ is a fluid model whose arrival rate at $t$ is given by $i$ if $S_{N(t)} = i$:

$$X(t) = S_{N(t)} = \sum_{n=0}^{\infty} S_n 1\{t_n \leq t < t_{n+1}\}.$$  

We assume that the random variable $T_n$ has distribution $F_T(\cdot)$, density $f_T(\cdot)$, and mean $\mu_T$ and the random variable $S_i$ has probability mass function $f_S(\cdot)$. Obviously, $X(t)$ is a stationary process with $E[X(t)] = E[S]$ and $Var[X(t)] = Var[S]$.

Now we consider the relationship between the ACF $\rho(\tau) = E[(X(t) - \mu_X)(X(t+\tau) - \mu_X)]/Var[X(t)]$ and the scene duration distribution $F_T(\cdot)$. Mandelbrot [76] gave an expression for the ACF for the discretely distributed duration of $T_i$. Here we give a simple and useful derivation different from that in [40, 76]. Consider two
source rates $X(t)$ and $X(t+\tau)$ at a distance of $\tau$ interval. Then, the auto-covariance of $X(t)$ is obtained as follows.

\[
\text{Cov}(X(t), X(t+\tau)) = E[(S_{N(t)} - \mu_X)^2] \Pr[N(t+\tau) = N(t)] + E[(S_{N(t)} - \mu_X)(S_{N(t+\tau)} - \mu_X)] \Pr[N(t+\tau) \neq N(t)]
\]
\[
= \text{Var}(S_{N(t)}) \int_{\tau}^{\infty} \Pr[u \leq T_{N(t)} < udu, t + \tau \leq t_{N(t)} + u]du
\]
\[
= \text{Var}(S) \int_{\tau}^{\infty} \frac{u \cdot f_T(u) u - \tau}{\mu_T} du
\]
\[
= \text{Var}(S) \left( 1 - \frac{\tau}{\mu_T} + \frac{1}{\mu_T} \int_{\tau}^{\infty} (\tau - u) f_T(u)du \right).
\]

The first equality is just to partition the events into \{\text{$N(t) = N(t+\tau)$}\} and \{\text{$N(t) \neq N(t+\tau)$}\}. In the second equality, the term for \{\text{$N(t) \neq N(t+\tau)$}\} becomes zero because of independence of $S_n$. In the third equation, we use the key renewal theorem [105] for the integrand and the property of identical distribution of $S_n$. We finally obtain the following relationship between the ACF and scene duration in the SL process. 

\[
\rho(t) = 1 - \frac{t}{\mu_T} + \frac{1}{\mu_T} \int_{0}^{t} (t - \tau) f_T(\tau)d\tau. \quad (3.2)
\]

Differentiation of the equation yields simple relations between the density of $T$ and the ACF.

\[
f_T(t) = \mu_T \rho''(t). \quad (3.3)
\]

We also present their discrete-time version without proof:

\[
\rho(n) = 1 - \frac{n}{\mu_T} + \frac{1}{\mu_T} \sum_{k=1}^{n} (n - k) f_T(k), \quad (3.4)
\]

and

\[
f_T(k) = \mu_T \nabla_2 \rho(k), \quad (3.5)
\]

where $\nabla_2 \rho(k) = \rho(k - 1) - 2\rho(k) + \rho(k + 1)$.
Table 3.1: Statistics for the traces used in our experiment (cells/frame or cells/GOP)

<table>
<thead>
<tr>
<th>trace</th>
<th>mean</th>
<th>std</th>
<th>coef. of var</th>
<th>peak</th>
<th>peak/mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>intra-coded <em>Star Wars</em></td>
<td>578.9</td>
<td>130.2</td>
<td>0.23</td>
<td>1,634.6</td>
<td>2.82</td>
</tr>
<tr>
<td>VIC Trace A</td>
<td>85.2</td>
<td>61.3</td>
<td>0.72</td>
<td>433.5</td>
<td>5.09</td>
</tr>
<tr>
<td>VIC Trace B</td>
<td>71.9</td>
<td>39.0</td>
<td>0.82</td>
<td>544.3</td>
<td>7.57</td>
</tr>
<tr>
<td>MPEG-GOP smoothed</td>
<td>589.1</td>
<td>331.6</td>
<td>0.56</td>
<td>3,796.7</td>
<td>6.44</td>
</tr>
</tbody>
</table>

3.2.2 The SL process as a Video Traffic Model

Now we consider application of the SL process to video traffic modeling. In general, the traffic pattern of a coded video trace depends on both the inherent variation of the visual information and the coding algorithm used for data compression. Especially, an MPEG coder generates periodic burstiness due to its picture-dependent coding algorithm. Here we focus on the inherent characteristics in VBR video traffic. So we use intra-coded *Star Wars* video traffic data [38] instead of MPEG traffic in order not to consider the side-effects of video coding. Table 3.1 shows the basic statistics for the traces used in our experiments, JPEG-coded *Star Wars* [37], VIC (H.H61) Trace A, VIC Trace B [104], and GOP smoothed and concatenated MPEG trace [98]. Here we assume 48 Bytes payload in 53 Bytes ATM cell size.

In our application, the scene size, i.e., the source rate in a scene, corresponds to $S_i$ and the scene duration corresponds to $T_i$. In general, the scene size process might not be i.i.d.; the real correlation depends on the definition of scene. In [4] and [54], the correlation is modeled by a discrete Markov chain. However, the large number of parameters in the transition matrix of semi-Markovian processes makes the matching procedure difficult. Furthermore, the performance estimation is very sensitive to its state definition [54]. This is because semi-Markovian models cannot
Figure 3.1: A sample path for the SLCC with the same ACF and the marginal distribution as the intra-coded *Star Wars* trace

capture the rate-distribution (first-order statistics) of the empirical traces with a limited number of states.

On the other hand, in the SL process assuming that the scene size process is also a renewal process, the SL process has the property that the marginal distribution and the ACF are determined solely by $S_i$ and $T_i$, respectively. Thus, we can easily match the rate-distribution and ACF of the model to those of the empirical traces. Furthermore, as we will show, our modeling based on the SL process requires only five parameters for VBR video traffic.

### 3.2.3 The SLCC and the C-DAR(1) Model

Figure 3.2 shows that the empirical ACF of video traffic seems to be a kind of exponential function (SRD) in the small lag region, and a kind of hyperbolic function (LRD) in the large lag region. Thus, we consider the *shifting-level process with a*
compound correlation \((SLCC)\) of the exponential and the hyperbolic as follows.

\[
\rho(t) = \begin{cases} 
\rho_e(t) = e^{-t/\tau}, & \text{for } 0 < t < t_0, \\
\rho_h(t) = c_0(t + t_1)^{-\beta}, & \text{for } t_0 < t
\end{cases}
\] (3.6)

First, we obtain

\[
\rho''_e(t) = (-e^{-t/\tau}/\tau)' = e^{-t/\tau}/\tau^2
\] (3.7)

\[
\rho''_h(t) = c_0(-\beta (t + t_1)^{-(\beta+1)})' = c_0(\beta(\beta + 1) (t + t_1)^{-1}(\beta+2)).
\] (3.8)

Then we find that \(T\) must also follow the exponential and hyperbolic distribution.

Figure 3.1 shows a sample trace of the SLCC process whose model parameters are matched with a Star Wars trace [38] using the parameter matching algorithm in Section 3.3.

Here we consider the relationship between the SLCC process and the well-known DAR(1) model [48]. The DAR(1) model is a Markov chain with transition matrix

\[
P = \rho I + (1 - \rho)Q,
\] (3.9)

where \(\rho\) is the autocorrelation and \(Q\) is a state transition matrix with identical rows equal to the marginal distribution.

To compare the DAR(1) model with the SL process, which is a continuous-time model, we consider the continuous-time analog (C-DAR(1)) of the DAR(1) model, which was named and studied in [115]. The motivation of the DAR(1) model is not discrete-time modeling but accurate modeling of the marginal distribution in an empirical trace [48, 115]; the original AR(1) model follows a Gaussian distribution. Furthermore, the exact match of DAR(1) and C-DAR(1) is verified in [115]. The C-DAR(1) model is a kind of SLCC process with exponential autocorrelation only in the SLCC and thus \(\rho(t) = e^{-t/\tau}\).

The C-DAR(1) model is a SRD process \((\sum \rho(t) < \infty)\), and cannot capture the heavy-tail properties in ACF of real video traffics. On the other hand, the SLCC is
a LRD process ($\sum \rho(t) = \infty$). Since the SLCC and the C-DAR(1) model have the same rate-distribution and short-term correlation structure except the hyperbolic tail, comparing the queueing performances of them will reveal the effects of long-range dependence in video traffic.

### 3.3 Parameter Matching of the SL Process

#### 3.3.1 The Proposed Parameter Matching Algorithm

Now we consider an algorithm which matches the parameters of the SLCC to the statistics of a real video traffic. In the SL process, the histogram is determined by the scene size process, $S_t$, and the ACF is determined by the scene duration process, $T_t$.

First, we propose the matching procedure for the ACF of the SLCC process. To approximate the ACF of a real video sequence with two mathematical functions, we have to determine the values of 5 parameters $\tau$, $\beta$, $t_0$, $t_1$, and $c_0$. We obtain the exponential decaying rate $\tau$ using

$$\tau = -t/\ln(\rho(t)),$$

(3.10)

where we use $t = 10$ frames, and the hyperbolic decaying rate

$$\beta = 2 - 2H$$

(3.11)

by the Hurst parameter estimation [15]. For parameters $c_0$ and $t_0$, we further assume that the ACF and the cumulative distribution function (CDF) of distribution are continuous. Then, we obtain the two relations:

$$\rho_e(t) = \rho_h(t)$$

$$\rho_e'(t) = \rho_h'(t)$$
at \( t = t_0 \). That is,

\[
e^{-t_0/\tau} = c_0(t_0 + t_1)^{-\beta}
\]

\[
\frac{1}{\tau} e^{-t_0/\tau} = c_0(-\beta)(t_0 + t_1)^{-(\beta+1)}
\]

Solving the system of equations, we finally obtain

\[
t_0 = \beta\tau - t_1
\]

\[
c_0 = (\tau\beta)^\beta e^{-(\beta-t_1/\tau)}
\]

Because \( t_0 \) and \( c_0 \) are functions of \( t_1 \), we use a least square fitting to determine the parameter \( t_1 \). Furthermore, from the relation \( f_T(t) = \mu_T \rho''(t) \), and \( \int_0^\infty f_T(t)dt = 1 \), we obtain \( \mu_T = \tau \).

For the random number generation, we here obtained the CDF for the compound ACF model.

\[
F_T(t) = 1 + \mu_T \rho'(t) = \begin{cases} 
1 - e^{-t/\tau}, & \text{for } t < t_0, \\
1 - \tau \beta c_0 (t + t_1)^{-(\beta+1)}, & \text{for } t_0 \leq t 
\end{cases}
\]

and the inverse function of it

\[
F_T^{-1}(y) = \begin{cases} 
-\tau \ln(1-y), & \text{for } y < 1 - e^{-\beta+t_1/\tau} \\
\left( \frac{\beta\tau}{1-y} \right)^{1/(\beta+1)} - t_1, & \text{for } y \geq 1 - e^{-\beta+t_1/\tau}.
\end{cases}
\]

The probability mass function for frame size \( f_S(i) \) is modeled by the marginal distribution of an empirical trace, \( (f_S(0), f_S(1), \ldots, f_S(M-1), f_S(M)) \), where \( f_S(i) \) is the negative binomial distribution

\[
f_S(i) = \binom{-\tau}{i} p^r(-q)^i = \binom{i+r-1}{i} p^r q^i, \quad i = 0, 1, 2, \ldots, N
\]

and \( f_S(M) = 1 - \sum_{i<M} f_S(i) \), and \( M \) is the peak rate in cells per frame [38, 48].

The mean and variance of this distribution are

\[
E[X(t)] = \frac{\tau(1-p)}{p} \quad \text{and} \quad \text{Var}[X(t)] = \frac{\tau(1-p)}{p^2},
\]
respectively. Here, $0 < p < 1$, $q = 1 - p$, and $r > 0$. Thus, the parameters are obtained by

$$p = \frac{E[X(t)]}{\text{Var}[X(t)]} \quad \text{and} \quad r = \frac{E[X(t)]^2}{\text{Var}[X(t)] - E[X(t)]}.$$  \hspace{1cm} (3.20)

For a multiplexed stream, we still assume the negative binomial distribution and the auto-covariance function. It can be easily shown that the auto-covariance function of homogeneous function is exactly same as that of a single source. For distribution of multiplexed sources, we will check our assumption through examining the probability distributions from real histogram of multiplexed empirical traces and from the mathematical models. Instead of the well-known Q-Q plot [66], we will examine the values where its negative cumulative distribution reaches several specific probabilities. This is because the Q-Q plot technique visualizes closeness between two distributions, it does not provide quantitative comparison.

First, we define a random variable $Y^j(N) = \frac{1}{N} \sum_{i=1}^{N} X_i$, where $N$ and $j$ denote the number of multiplexed sources and the distribution used (i.e., real distribution, Gaussian, Nbin), respectively. Then, we compare the values of normalized deviation $k^j(N, p) = (y^j - E[Y^j(N)])/E[Y^j(N)]$ such that $p = \text{Prob}[Y^j(N) > y]$. The normalized deviation represents the value of random variable where its negative cumulative probability reach a specific probability $p$. Note that the normalization is done by mean value instead of standard deviation. This is because the value of $1/(k^j(N, p) + 1)$ is the link utilization at bufferless multiplexing system for the given cell loss probability guarantee. In other word, $k^j(N, p) \cdot E(Y^j(N)$ is the extra bandwidth to reserve for guaranteeing cell loss ratio than $p$ in bufferless model. Numerical results to verify our assumption will be given in the following numerical section.

Here we summarize the parameter matching procedure:

For $F_T$,
3.3 Parameter Matching of the SL Process

- Obtain \( \tau \) by equation (3.10), and \( \beta \) by equation (3.11).

- Obtain \( t_0 \), \( c_0 \) and \( t_1 \) from the least square fitting and the relationship (3.14) and (3.15).

For \( f_S \),

- Obtain \( p, r \) by equations (3.20).

3.3.2 Numerical Results for Empirical Traces

The parameters of the SLCC is obtained as follows. First, since the derivations in the previous sections are continuous-time versions, we round off scene durations into integer numbers. First, for the JPEG coded Star Wars traces (Fig. 3.2): for the ACF, \( \tau = 82.83 \) frames, \( \beta = 0.39 \), \( t_0 = 30.13 \) frames, \( t_1 = 3.0 \) frames, and \( c_0 = 2.82 \), and for the histogram, \( M = 1,634 \) cells, \( E[X(t)] = 578.9 \) cells/frame, and \( \sqrt{\text{Var}[X(t)]} = 130.2 \) and thus \( p = 0.0341 \) and \( r = 20.44 \sim 20 \) are obtained. In the same way, we obtained the parameters for various traces, vic trace A, vic trace B, and GOP-smoothed MPEG the same way, and verified the performance of our modeling (Figs. 3.4): For the vic trace A, \( t_0 = 12.4361 \) frames, \( t_1 = 1 \) frames, \( \beta = 0.63 \), \( c_0 = 2.8679 \), \( a = 0.9542 \), and \( \tau = 21.3272 \) frames. For the vic trace B, \( t_0 = 3.2950 \) frames, \( t_1 = 10 \) frames, \( \beta = 0.3500 \), \( c_0 = 2.2679 \), \( a = 0.9740 \), and \( \tau = 37.9856 \) frames. For the GOP smoothed MPEG, \( t_0 = 1.33 \) GOPs, \( t_1 = 1 \) GOPs, \( \beta = 0.26 \), \( c_0 = 1.0741 \), \( a = 0.8944 \) GOPs, and \( \tau = 8.9615 \) GOPs (with a GOP size of 12 frames).

For the C-DAR(1) model, we use the same value of \( \tau \) (for the exponential decay rate), and \( p \) and \( r \) (for the histogram) as for the SLCC process.

Now we examine how accurately the SLCC model emulates the ACF of the empirical traces. Figure 3.2 shows four ACFs: (a) original, (b) exponential, (c)
Figure 3.2: The ACFs for original, exponential, hyperbolic and SLCC (JPEG)

Figure 3.3: The ACFs for original, exponential, hyperbolic and SLCC (GOP smoothed MPEG)
Figure 3.4: The ACFs for original, exponential, hyperbolic and SLCC (VIC Trace A, VIC trace B)
hyperbolic, and (d) generated SLCC. The generated SLCC from the proposed parameter matching algorithm provides a very good fit at both the small and large lags, while the exponential curve of the C-DAR(1) underestimates in the region of large lags. Therefore, with the assumption that binomial distribution accurately matches the frame size distribution, we can conclude that the statistical characteristics of the SL process are equivalent to those of the original video trace up to second-order statistics.

We also verify the accuracy in approximation of empirical distribution with the negative binomial distribution. As we mentioned the previous section, we will investigates not only the distribution of single input but also that of multiplexed source. We use the normalized deviation \(k_j(N,p)\) as measure of closeness, where \(j, N, p\) denotes the distributions, the number of multiplexed sources, and the negative cumulative probabilities. Figure 3.5 show the accuracy increases with the number of sources, and the negative binomial approximation outperforms the Gaussian approximation. And yet with more than 100 sources, the difference is not significant, which advocates the central limit theorem based approach [22] at central nodes in core network. Also we provides experiment results for the GOP-smoothed MPEG, VIC trace A, and B.

3.4 An Efficient Analysis of the SL/D/1/K Queueing System

3.4.1 The Quantization Reduction Method

In this section, we present an efficient analysis method for the SL/D/1/K queueing system. We consider a single server fluid queueing system with a buffer of size \(K\) cells and a deterministic service rate of \(C\) cells per frame. Input arrives at rate \(i\),
Table 3.2: Normalized deviations for negative cumulative probabilities: Real trace, Negative Binomial and Gaussian approximation

**JPEG coded Star Wars**

<table>
<thead>
<tr>
<th>Prob.</th>
<th>NoS</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0E-1</td>
<td>Real</td>
<td>0.28</td>
<td>0.20</td>
<td>0.14</td>
<td>0.10</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>0.29</td>
<td>0.20</td>
<td>0.14</td>
<td>0.10</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>0.28</td>
<td>0.20</td>
<td>0.14</td>
<td>0.10</td>
<td>0.07</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>1.0E-3</td>
<td>Real</td>
<td>1.10</td>
<td>0.64</td>
<td>0.41</td>
<td>0.27</td>
<td>0.18</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>0.84</td>
<td>0.56</td>
<td>0.38</td>
<td>0.26</td>
<td>0.18</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>0.69</td>
<td>0.49</td>
<td>0.34</td>
<td>0.24</td>
<td>0.17</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>1.0E-5</td>
<td>Real</td>
<td>1.82</td>
<td>1.05</td>
<td>0.64</td>
<td>0.41</td>
<td>0.27</td>
<td>0.18</td>
<td>0.12</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>1.26</td>
<td>0.82</td>
<td>0.55</td>
<td>0.37</td>
<td>0.25</td>
<td>0.17</td>
<td>0.12</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>0.95</td>
<td>0.67</td>
<td>0.47</td>
<td>0.33</td>
<td>0.23</td>
<td>0.16</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>1.0E-7</td>
<td>Real</td>
<td>1.82</td>
<td>1.22</td>
<td>0.75</td>
<td>0.47</td>
<td>0.31</td>
<td>0.20</td>
<td>0.14</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>1.64</td>
<td>1.06</td>
<td>0.69</td>
<td>0.47</td>
<td>0.32</td>
<td>0.22</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>1.17</td>
<td>0.82</td>
<td>0.58</td>
<td>0.41</td>
<td>0.29</td>
<td>0.20</td>
<td>0.14</td>
<td>0.10</td>
</tr>
</tbody>
</table>

**GOP-smoothed MPEG**

<table>
<thead>
<tr>
<th>Prob.</th>
<th>NoS</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0E-1</td>
<td>Real</td>
<td>0.73</td>
<td>0.52</td>
<td>0.36</td>
<td>0.26</td>
<td>0.18</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>0.75</td>
<td>0.53</td>
<td>0.37</td>
<td>0.26</td>
<td>0.18</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>0.72</td>
<td>0.51</td>
<td>0.36</td>
<td>0.25</td>
<td>0.18</td>
<td>0.12</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>1.0E-3</td>
<td>Real</td>
<td>3.39</td>
<td>2.04</td>
<td>1.25</td>
<td>0.79</td>
<td>0.52</td>
<td>0.34</td>
<td>0.23</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>2.65</td>
<td>1.68</td>
<td>1.10</td>
<td>0.72</td>
<td>0.49</td>
<td>0.33</td>
<td>0.23</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>1.73</td>
<td>1.23</td>
<td>0.86</td>
<td>0.61</td>
<td>0.43</td>
<td>0.30</td>
<td>0.21</td>
<td>0.15</td>
</tr>
<tr>
<td>1.0E-5</td>
<td>Real</td>
<td>5.45</td>
<td>3.23</td>
<td>2.00</td>
<td>1.24</td>
<td>0.78</td>
<td>0.51</td>
<td>0.34</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>4.35</td>
<td>2.66</td>
<td>1.68</td>
<td>1.08</td>
<td>0.71</td>
<td>0.48</td>
<td>0.32</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>2.40</td>
<td>1.69</td>
<td>1.20</td>
<td>0.84</td>
<td>0.60</td>
<td>0.42</td>
<td>0.30</td>
<td>0.21</td>
</tr>
<tr>
<td>1.0E-7</td>
<td>Real</td>
<td>5.45</td>
<td>3.88</td>
<td>2.36</td>
<td>1.44</td>
<td>0.91</td>
<td>0.58</td>
<td>0.38</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>5.98</td>
<td>3.57</td>
<td>2.20</td>
<td>1.39</td>
<td>0.91</td>
<td>0.60</td>
<td>0.40</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>2.92</td>
<td>2.06</td>
<td>1.46</td>
<td>1.03</td>
<td>0.73</td>
<td>0.51</td>
<td>0.36</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Table 3.3: More Numerical results on normalized deviations

**VIC Trace A**

<table>
<thead>
<tr>
<th>Prob.</th>
<th>NoS</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0E-1</td>
<td>Real</td>
<td>1.03</td>
<td>0.69</td>
<td>0.47</td>
<td>0.33</td>
<td>0.23</td>
<td>0.16</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>0.96</td>
<td>0.68</td>
<td>0.48</td>
<td>0.33</td>
<td>0.23</td>
<td>0.16</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>0.92</td>
<td>0.65</td>
<td>0.46</td>
<td>0.32</td>
<td>0.23</td>
<td>0.16</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>1.0E-3</td>
<td>Real</td>
<td>2.85</td>
<td>1.91</td>
<td>1.30</td>
<td>0.88</td>
<td>0.60</td>
<td>0.41</td>
<td>0.29</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>3.72</td>
<td>2.32</td>
<td>1.49</td>
<td>0.97</td>
<td>0.65</td>
<td>0.44</td>
<td>0.30</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>2.23</td>
<td>1.5e</td>
<td>1.12</td>
<td>0.78</td>
<td>0.55</td>
<td>0.39</td>
<td>0.27</td>
<td>0.19</td>
</tr>
<tr>
<td>1.0E-5</td>
<td>Real</td>
<td>4.09</td>
<td>2.78</td>
<td>1.88</td>
<td>1.27</td>
<td>0.86</td>
<td>0.59</td>
<td>0.41</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>6.31</td>
<td>3.77</td>
<td>2.33</td>
<td>1.49</td>
<td>0.96</td>
<td>0.64</td>
<td>0.43</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>3.08</td>
<td>2.18</td>
<td>1.54</td>
<td>1.09</td>
<td>0.77</td>
<td>0.54</td>
<td>0.38</td>
<td>0.27</td>
</tr>
<tr>
<td>1.0E-7</td>
<td>Real</td>
<td>4.09</td>
<td>3.16</td>
<td>2.12</td>
<td>1.44</td>
<td>0.97</td>
<td>0.66</td>
<td>0.46</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>8.85</td>
<td>5.15</td>
<td>3.10</td>
<td>1.93</td>
<td>1.23</td>
<td>0.81</td>
<td>0.54</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>3.75</td>
<td>2.65</td>
<td>1.87</td>
<td>1.33</td>
<td>0.93</td>
<td>0.66</td>
<td>0.46</td>
<td>0.33</td>
</tr>
</tbody>
</table>

**VIC Trace B**

<table>
<thead>
<tr>
<th>Prob.</th>
<th>NoS</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0E-1</td>
<td>Real</td>
<td>1.18</td>
<td>0.78</td>
<td>0.54</td>
<td>0.38</td>
<td>0.26</td>
<td>0.18</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>1.09</td>
<td>0.78</td>
<td>0.55</td>
<td>0.38</td>
<td>0.27</td>
<td>0.19</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>1.05</td>
<td>0.74</td>
<td>0.52</td>
<td>0.37</td>
<td>0.26</td>
<td>0.18</td>
<td>0.13</td>
<td>0.09</td>
</tr>
<tr>
<td>1.0E-3</td>
<td>Real</td>
<td>3.33</td>
<td>2.23</td>
<td>1.50</td>
<td>1.02</td>
<td>0.69</td>
<td>0.48</td>
<td>0.33</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>4.48</td>
<td>2.77</td>
<td>1.76</td>
<td>1.14</td>
<td>0.75</td>
<td>0.51</td>
<td>0.34</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>2.55</td>
<td>1.80</td>
<td>1.27</td>
<td>0.90</td>
<td>0.63</td>
<td>0.45</td>
<td>0.31</td>
<td>0.22</td>
</tr>
<tr>
<td>1.0E-5</td>
<td>Real</td>
<td>6.56</td>
<td>3.65</td>
<td>2.27</td>
<td>1.48</td>
<td>0.99</td>
<td>0.68</td>
<td>0.46</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>7.74</td>
<td>4.57</td>
<td>2.79</td>
<td>1.75</td>
<td>1.13</td>
<td>0.74</td>
<td>0.50</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>3.51</td>
<td>2.48</td>
<td>1.76</td>
<td>1.24</td>
<td>0.87</td>
<td>0.62</td>
<td>0.43</td>
<td>0.31</td>
</tr>
<tr>
<td>1.0E-7</td>
<td>Real</td>
<td>6.56</td>
<td>4.25</td>
<td>2.63</td>
<td>1.69</td>
<td>1.13</td>
<td>0.76</td>
<td>0.52</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Nbin</td>
<td>8.09</td>
<td>6.31</td>
<td>3.75</td>
<td>2.30</td>
<td>1.46</td>
<td>0.94</td>
<td>0.63</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>Gauss</td>
<td>4.28</td>
<td>3.03</td>
<td>2.14</td>
<td>1.51</td>
<td>1.07</td>
<td>0.75</td>
<td>0.53</td>
<td>0.37</td>
</tr>
</tbody>
</table>
when the SLCC process \( X(t) \) is state \( i \) at time \( t \). Let \( L(t) \) be the number of cells in the system at time \( t \). We are interested in the stochastic process \( \{X(t), L(t), t \geq 0\} \), which characterizes the dynamics of the system. Let \( t_n \) is the state transition epoch of the SL process. Then \( \{(X_n, L_n) \equiv (X(t_n^+), L(t_n^+))\} \) is an embedded Markov chain. However, the exact analysis of this system is very difficult, because \( L_n \) takes a continuous value in \([0, K]\) and the sojourn times of states in a SL process do not follow an exponential distribution. Therefore, we present an approximation method, which we call quantization reduction method, to approximate the system behaviors. The idea of the state space quantization method is to reduce the state space of a continuous queue size \( L_n, [0, K] \) to a finite set of quantization points \( \{0, h, 2h, \cdots, Dh = K\} \), where \( h \) is the quantization step and \( D + 1 \) is the number of the queue states.

First, we define 3 auxiliary processes \( L_a(t), L_t(t) \) and \( L^I(t) \) for \( L(t) \). The su-
perscripts \(u, a,\) and \(l\) mean upper bound, approximation, and lower bound, respectively. At every transition epoch of input, \(t_n\), the processes, \(L^u(t),\) \(L^a(t),\) and \(L^l(t)\) are approximated by their respective quantizers: 
\[L^u_n \equiv L^u(t^+_{n}) = \mathcal{F}^u L^u(t^-_{n}),\]
\[L^a_n \equiv L^a(t^+_{n}) = \mathcal{F}^a L^a(t^-_{n}),\]
and 
\[L^l_n \equiv L^l(t^+_{n}) = \mathcal{F}^l L^l(t^-_{n}).\] These quantization functions are all staircase functions (Figure 3.6) and given as follows.

\[
\mathcal{F}^a := \begin{cases} 
0, & l = 0, \\
\left(lh - \frac{h}{2}, lh + \frac{h}{2}\right) & l = 1, 2, \ldots D - 1, \\
\left[Dh - \frac{h}{2}, K\right] & l = D.
\end{cases} 
\] (3.21)

\[
\mathcal{F}^u := \begin{cases} 
0 \to 0, & l = 0, \\
\left((l - 1)h, lh\right] \to lh, & l = 1, 2, \ldots D.
\end{cases} 
\] (3.22)

\[
\mathcal{F}^l := \begin{cases} 
\left[lh, (l + 1)h\right) \to lh, & l = 0, 1, 2, \ldots D - 1, \\
\left[(D - 1)h, K\right] \to D, & l = D.
\end{cases} 
\] (3.23)
3.4 An Efficient Analysis of the SL/D/1/K Queueing System

Figure 3.7: Sample paths for the $L(t)$, $L^a(t)$, $L^u(t)$, and $L^l(t)$.

Figure 3.7 illustrates a sample path for each stochastic process, $L^u_n$, $L^a_n$, $L^l_n$, and $L_n$, and $L^u_n(t)$, $L^a_n(t)$, $L^l_n(t)$, and $L_n(t)$, to show the relationship among them. Since $L^l_0 \leq L^a_0 \sim L_0 \leq L^u_0$ and 4 systems are loaded by the same input, it is clear that $L^l_n \leq L^a_n \sim L_n \leq L^u_n, i = 1, 2, 3, \ldots$. And, it is also clear that $L^l(t) \leq L^a(t) \sim L(t) \leq L^u(t)$.

The processes $\{X_n, L^a_n\}$, $\{X_n, L^u_n\}$, and $\{X_n, L^l_n\}$ are all irreducible and positive recurrent embedded Markov chains and thus the queuing analysis is based on the transition matrix of the embedded processes. Let $P^a$, $P^u$, and $P^l$ be the one-step transition probability matrix of $\{(X_n, L^a_n)\}$, $\{(X_n, L^u_n)\}$, and $\{(X_n, L^a_n)\}$, respectively. The elements of the transition matrices are defined by

$$p^e_{(i,l),(j,k)} \equiv Pr\{(X_{n+1}, L^e_{n+1}) = (j,k)|(X_n, L^e_n) = (i,l)\}, c = u, a, l. \quad (3.24)$$

Noting that during $[t_n, t_{n+1})$, input arrives at uniform rate of $i$ cells per frame, the elements of each transition matrix is obtained as follows, respectively.
\[ P_{(i,l),(j,k)}^S = \begin{cases} 
0, & \text{for } i \geq C, l > k, \\
fs(j)(FT\left(\frac{(k-l+\frac{1}{2})h}{(i-C)}\right)) - FT\left(\frac{(k-l-\frac{1}{2})h}{(i-C)}\right)^+, & \text{for } i \geq C, l \leq k, \\
fs(j)(FT\left(\frac{(l-k+\frac{1}{2})h}{(C-\bar{C})}\right)) - FT\left(\frac{(l-k-\frac{1}{2})h}{(C-\bar{C})}\right)^+, & \text{for } i > C, l \geq k, \\
0, & \text{for } i < C, l > k. 
\end{cases} \tag{3.25} \]

\[ P_{(i,l),(j,k)}^U = \begin{cases} 
0, & \text{for } i \geq C, l \geq k, \\
fs(j)(FT\left(\frac{(k-l)h}{(i-C)}\right)) - FT\left(\frac{(k-l-1)h}{(i-C)}\right)^+, & \text{for } i \geq C, l < k, \\
fs(j)(FT\left(\frac{(l-k+1)h}{C-i}\right)) - FT\left(\frac{(l-k)h}{C-i}\right)^+, & \text{for } i < C, l \geq k, \\
0, & \text{for } i < C, l < k. 
\end{cases} \tag{3.26} \]

\[ P_{(i,l),(j,k)}^I = \begin{cases} 
0, & \text{for } i \geq C, l > k, \\
fs(j)(FT\left(\frac{(k-l+1)h}{(i-C)}\right)) - FT\left(\frac{(k-l)h}{(i-C)}\right)^+, & \text{for } i \geq C, l \leq k, \\
fs(j)(FT\left(\frac{(l-k)h}{C-i}\right)) - FT\left(\frac{(l-k-1)h}{C-i}\right)^+, & \text{for } i < C, l > k, \\
0, & \text{for } i < C, l \leq k. 
\end{cases} \tag{3.27} \]

We define the limiting probabilities \( \pi_{i,l}^c = \lim_{n \to \infty} Pr\{X_n^c = i, L_n = l\}, c = u, a, l \). Then the limiting probability vector \( \pi^c = (\pi_{0,0}^c, \pi_{0,1}^c, \ldots, \pi_{0,D}^c, \pi_{1,0}^c, \ldots, \pi_{M,D}^c) \) is uniquely determined by

\[ \pi^c P^c = \pi^c \text{ and } \pi^c e^T = 1, c = u, a, l, \tag{3.28} \]

where \( e \) is a unity vector of dimension \((D + 1)(M + 1)\).

The computation can be obtained using standard numerical methods for solving system of equations. However, substantial saving in computation can be obtained if the renewal property of the SL process is used. First, we obtain a new transition matrix \( \hat{P}^c, c = u, a, l \) by merging the elements of the original transition matrix \( P^c \),

\[ \hat{P}_{i,k}^c = \sum_{i,j} P_{(i,l),(j,k)}^c \tag{3.29} \]
and then solve the system of equations,

\[ \hat{\pi}^c \hat{P}^c = \hat{\pi}^c \quad \text{and} \quad \hat{\pi}^c \mathbf{e}^T = 1, c = u, a, l, \]

where \( \mathbf{e} \) is a unitary vector of dimension \( D + 1 \). Finally, the steady-state probability \( \pi^c \) is obtained by multiplying the probability of rates,

\[ \pi_{i, l}^c = f_S(i) \cdot \hat{\pi}^c, c = u, a, l. \]

This observation reduces the dimension of the matrix from \( (D + 1)(M + 1) \times (D + 1)(M + 1) \) into \( (D + 1) \times (D + 1) \).

We now obtain the survival function, i.e., the complementary queue distribution at arbitrary times, and cell loss probability as performance measures. Since \( L^t(t) \leq L^u(t) \leq L^u(t) \), it is also clear that \( Pr[L^u(t) > x] \leq Pr[L^a(t) > x] \leq Pr[L^t(t) > x] \) for \( x \geq 0 \) and \( CLP^d \leq CLP^u \sim CLP^c \leq CLP^u \). The survival function \( Pr[L(t) > x] \) is obtained by calculating the fraction of time when the \( L(t) \) stays above system size \( x \) between two successive embedded points, \( t_n \) and \( t_{n+1} \):

\[
Pr[L^c(t) > x] = \frac{\sum_i \sum_l \pi_{i, l}^c E(T > x | (i, l))}{E(T)}, \quad c = u, a, l
\]

where the conditional expectation is obtained by

\[
E(T > x | (i, l)) = \begin{cases} 
E(T) & \text{for } i \geq C, l \geq x \\
E((T - \frac{x - hl}{i - c})^+ | (i, l)) & \text{for } i \geq C, l < x \\
E((\frac{hl - x}{C - i} - T)^+ | (i, l)) & \text{for } i < C, l > x \\
0 & \text{for } i < C, l \leq x
\end{cases}
\]

The cell loss probabilities \( CLP^c, c = u, a, l \) are defined by the fraction of overflow data among the total arrivals. Noting that overflow occurs only when the input rate is larger than \( C \), we obtain

\[
CLP^c \equiv \frac{\sum_{i > C} \sum_l \pi_{i, l}^c (i - C) E((T - \frac{K - hl}{i - c})^+ | (i, l))}{E(T)E(S)}, \quad c = u, a, l.
\]
3.4.2 The Efficiency of the Quantization Reduction Method

The accuracy of the quantization reduction analysis depends on the input traffic: roughly speaking, as the product of its mean sojourn time and mean arrival rate increases, the accuracy also increases for the fixed quantization value. To verify the accuracy in video traffic application, we show numerical results for queue occupancy distributions and cell loss probabilities, varying the traffic intensity $U = 0.8$ (heavy load), $U = 0.6$ (moderate load), and $U = 0.4$ (light load).

In Table 3.4, 3.5, and 3.6, we show the queue distributions obtained by $L^u(t)$, $L^a(t)$, and $L^l(t)$ with system size $K = 50,000$ cells and two quantization values $h_1 = 5000$, $h_2 = 500$ cells. At all three traffic intensities, the errors $\epsilon = Pr[L^u > x] - Pr[L^l > x]$ are less than 1% for $h_1$ and 10% for $h_2$. Furthermore, the relative magnitude of error decreases at larger queue sizes.

Again we show the cell loss ratio for $U = 0.8, 0.6, 0.4$ and the same $h_1$ and $h_2$ as in the queue distribution in Tables 3.7, 3.8, and 3.9. The error bounds of cell loss ratio, $\epsilon = CLP^u - CLP^l$, are tighter than those for queue occupancy. It can be expected from that quantization reduction method estimates more accurately at large queue sizes. From these results, we conclude that the quantization reduction method is efficient and accurate in the queueing performances for VBR video traffic. Therefore, in the next section, although we show only the result of $L^a$ with quantization value $h_2$, for simplicity, the results are obtained within accuracy of 1%.

The embedded Markov chain method has been widely used in the area of teletraffic research. In general the computation load of the embedded Markov chain approach is an order of $O((MK)^2)$, where M is the number of the input state and K is the size of the system space, since it is based on the method for solving the system of equations. The existing numerical methods to reduce the computation load can
be classified into two groups. One is to use the characteristics of a matrix such as
the matrix geometric approach [80] and the quasi-birth death (QBD) approach [69],
and the other is to reduce the number of input states [11]. The first approaches
do reduce the computational load, but still not enough for video traffic application.
The second approaches also reduce the computational load with deep insight into
the queueing system, but it is very risky because the accuracy of results often de-
pends on the system condition [11]. Therefore, the usage of mathematical methods
has been limited only to a small number of input states and small queue sizes. This
is more severe in the case of video traffic, because of the long-term correlation and
large data size. Thus most video traffic models such as the DAR(1) and Markov
renewal process models [73] are only used as simulation models for traffic engineers.

Our quantization reduction method has the important advantage that the range
of error of the approximation $L^a(t)$ from $L(t)$ is given by providing the the upper
and lower bounds from the auxiliary processes $L^u(t)$ and $L^l(t)$. And thus we can
efficiently use the memory size and computation time for performance measures
by adjusting the quantization step. Furthermore, in the SL process we do not
suffer from computational load since the dimension of the one-step transition matrix
$\hat{P}^c, c = u, a, l$ does not depend on the size of the state space for input process. Thus
we can accurately match the histogram (first-order statistics) of an input process
without a significant increase in the computational load.

3.5 Queueing Performance Results

Now we conduct several queueing experiments: for a single stream, and for sta-
tistically multiplexed streams. The purpose of these experiments is to compare the
queueing behaviors of the SLCC and the C-DAR(1) model with that of an empir-
Table 3.4: Queue occupancies for $L^u(t)$, $L^a(t)$, and $L^l(t)$ ($U = 0.8$)

<table>
<thead>
<tr>
<th>$x$ (Kcells)</th>
<th>$P_r[L^u(t) &gt; x]$</th>
<th>$P_r[L^a(t) &gt; x]$</th>
<th>$P_r[L^l(t) &gt; x]$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h_1$</td>
<td>$h_2$</td>
<td>$h_1$</td>
</tr>
<tr>
<td>0</td>
<td>2.9053E-1</td>
<td>2.0920E-1</td>
<td>1.9750E-1</td>
</tr>
<tr>
<td>50</td>
<td>2.7884E-2</td>
<td>2.6643E-2</td>
<td>2.6643E-2</td>
</tr>
</tbody>
</table>

Table 3.5: Queue occupancies for $L^u(t)$, $L^a(t)$, and $L^l(t)$ ($U = 0.6$)

<table>
<thead>
<tr>
<th>$x$ (Kcells)</th>
<th>$P_r[L^u(t) &gt; x]$</th>
<th>$P_r[L^a(t) &gt; x]$</th>
<th>$P_r[L^l(t) &gt; x]$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h_1$</td>
<td>$h_2$</td>
<td>$h_1$</td>
</tr>
<tr>
<td>10</td>
<td>3.5069E-3</td>
<td>3.3170E-3</td>
<td>3.2959E-3</td>
</tr>
<tr>
<td>20</td>
<td>2.4710E-3</td>
<td>2.3964E-3</td>
<td>2.3890E-3</td>
</tr>
<tr>
<td>30</td>
<td>1.9921E-3</td>
<td>1.9572E-3</td>
<td>1.9539E-3</td>
</tr>
<tr>
<td>40</td>
<td>1.6979E-3</td>
<td>1.6845E-3</td>
<td>1.6832E-3</td>
</tr>
<tr>
<td>50</td>
<td>1.4918E-3</td>
<td>1.4910E-3</td>
<td>1.4910E-3</td>
</tr>
</tbody>
</table>

Table 3.6: Queue occupancies for $L^u(t)$, $L^a(t)$, and $L^l(t)$ ($U = 0.4$)

<table>
<thead>
<tr>
<th>$x$ (Kcells)</th>
<th>$P_r[L^u(t) &gt; x]$</th>
<th>$P_r[L^a(t) &gt; x]$</th>
<th>$P_r[L^l(t) &gt; x]$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$h_1$</td>
<td>$h_2$</td>
<td>$h_1$</td>
</tr>
<tr>
<td>10</td>
<td>2.9731E-5</td>
<td>2.9221E-5</td>
<td>2.9184E-5</td>
</tr>
<tr>
<td>20</td>
<td>2.1787E-5</td>
<td>2.1564E-5</td>
<td>2.1585E-5</td>
</tr>
<tr>
<td>30</td>
<td>1.7895E-5</td>
<td>1.7779E-5</td>
<td>1.7813E-5</td>
</tr>
<tr>
<td>40</td>
<td>1.5489E-5</td>
<td>1.5435E-5</td>
<td>1.5461E-5</td>
</tr>
<tr>
<td>50</td>
<td>1.3884E-5</td>
<td>1.3884E-5</td>
<td>1.3884E-5</td>
</tr>
</tbody>
</table>
### 3.5 Queueing Performance Results

#### Table 3.7: Cell loss probabilities for $L^u(t)$, $L^a(t)$, and $L^l(t)$ ($U=0.8$)

<table>
<thead>
<tr>
<th>$x$ (Kcells)</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_1$</th>
<th>$h_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.9287E-2</td>
<td>1.9287E-2</td>
<td>1.9287E-2</td>
<td>1.9287E-2</td>
<td>1.9287E-2</td>
<td>1.9287E-2</td>
</tr>
<tr>
<td>20</td>
<td>7.6814E-3</td>
<td>7.0617E-3</td>
<td>7.0555E-3</td>
<td>7.0354E-3</td>
<td>6.9085E-3</td>
<td>7.0133E-3</td>
</tr>
<tr>
<td>30</td>
<td>6.3626E-3</td>
<td>5.9833E-3</td>
<td>5.9775E-3</td>
<td>5.9664E-3</td>
<td>5.8816E-3</td>
<td>5.9522E-3</td>
</tr>
<tr>
<td>40</td>
<td>5.5759E-3</td>
<td>5.3169E-3</td>
<td>5.3121E-3</td>
<td>5.3048E-3</td>
<td>5.2423E-3</td>
<td>5.2946E-3</td>
</tr>
<tr>
<td>50</td>
<td>5.0428E-3</td>
<td>4.8516E-3</td>
<td>4.8476E-3</td>
<td>4.8423E-3</td>
<td>4.7933E-3</td>
<td>4.8344E-3</td>
</tr>
</tbody>
</table>

#### Table 3.8: Cell loss probabilities for $L^u(t)$, $L^a(t)$, and $L^l(t)$ ($U=0.6$)

<table>
<thead>
<tr>
<th>$x$ (Kcells)</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_1$</th>
<th>$h_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.3163E-3</td>
<td>1.3163E-3</td>
<td>1.3163E-3</td>
<td>1.3163E-3</td>
<td>1.3163E-3</td>
<td>1.3163E-3</td>
</tr>
<tr>
<td>20</td>
<td>5.0984E-4</td>
<td>5.0918E-4</td>
<td>5.0915E-4</td>
<td>5.0914E-4</td>
<td>5.0886E-4</td>
<td>5.0910E-4</td>
</tr>
<tr>
<td>40</td>
<td>3.8874E-4</td>
<td>3.8850E-4</td>
<td>3.8848E-4</td>
<td>3.8848E-4</td>
<td>3.8836E-4</td>
<td>3.8846E-4</td>
</tr>
<tr>
<td>50</td>
<td>3.5598E-4</td>
<td>3.5580E-4</td>
<td>3.5578E-4</td>
<td>3.5578E-4</td>
<td>3.5569E-4</td>
<td>3.5577E-4</td>
</tr>
</tbody>
</table>

#### Table 3.9: Cell loss probabilities for $L^u(t)$, $L^a(t)$, and $L^l(t)$ ($U=0.4$)

<table>
<thead>
<tr>
<th>$x$ (Kcells)</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_1$</th>
<th>$h_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.1704E-5</td>
<td>1.1704E-5</td>
<td>1.1704E-5</td>
<td>1.1704E-5</td>
<td>1.1704E-5</td>
<td>1.1704E-5</td>
</tr>
<tr>
<td>10</td>
<td>5.0632E-6</td>
<td>5.0631E-6</td>
<td>5.0631E-6</td>
<td>5.0631E-6</td>
<td>5.0631E-6</td>
<td>5.0631E-6</td>
</tr>
<tr>
<td>20</td>
<td>3.8651E-6</td>
<td>3.8651E-6</td>
<td>3.8651E-6</td>
<td>3.8651E-6</td>
<td>3.8651E-6</td>
<td>3.8651E-6</td>
</tr>
<tr>
<td>30</td>
<td>3.2945E-6</td>
<td>3.2945E-6</td>
<td>3.2945E-6</td>
<td>3.2945E-6</td>
<td>3.2945E-6</td>
<td>3.2945E-6</td>
</tr>
<tr>
<td>40</td>
<td>2.9400E-6</td>
<td>2.9400E-6</td>
<td>2.9400E-6</td>
<td>2.9400E-6</td>
<td>2.9400E-6</td>
<td>2.9400E-6</td>
</tr>
<tr>
<td>50</td>
<td>2.6909E-6</td>
<td>2.6909E-6</td>
<td>2.6909E-6</td>
<td>2.6909E-6</td>
<td>2.6909E-6</td>
<td>2.6909E-6</td>
</tr>
</tbody>
</table>
ical trace, so that we determine under what condition LRD is crucial for queueing behavior.

First, we examine the queueing behavior for a single source varying the input traffic load. The queueing performances loaded with the SLCC and the C-DAR(1) model are obtained by the quantization reduction method ($h_2$). For simulation with the original trace, we start the trace at a random number of frames, and upon reaching the end of the trace, wrap each source around to the beginning, so all 171,000 frames are used. We conduct simulation using the iterative equation

$$L_{n+1} = \min(\max(L_n + X_n - C, 0), K).$$

A summary of queue size distributions and cell loss probabilities with a single source for traffic load ($U$) 0.8 (heavy load), 0.6 (moderate load), and 0.4 (light load) is shown in Figures 3.8 and 3.9, respectively. Note that in the region where the C-DAR(1) provides accurate prediction, the merit of SLCC process, i.e., the effects of LRD could not be observed. For the queue distribution, we use a buffer size of 50,000 cells. Detailed investigation of the figures shows that both the C-DAR(1) model and the SLCC provide acceptable prediction (in order of magnitude) in the small buffer region. Interestingly, the C-DAR(1) model significantly underestimates in the region of $U = 0.8$ and a large queue size, whereas the SL process provides accurate prediction consistently.

The difference in the queueing results at heavy and light traffic loads can be explained as follows. The queueing system buffers the input arrival stream and thus memorizes its correlation. At a heavy traffic load, the busy period is long enough for the queueing system to be effected by LRD of the input traffic. However, whenever the queue size hits the bottom, i.e., $L(t) = 0$, the queue state is reset and forgets the past correlation of the input stream. In contrast, The resetting event
3.5 Queueing Performance Results

Figure 3.8: Queue occupancies for the empirical, SLCC, and C-DAR(1) inputs

Figure 3.9: Cell loss probabilities for the empirical, SLCC, and C-DAR(1) inputs
occurs more frequently as the traffic load decreases. Thus, at light traffic load long term correlation in input traffic does not have a significant impact on the queuing behavior. A similar argument can be found in [49].

We now consider the effects of multiplexing independent sources. In the SL model, the parameter matching and queueing analysis for the superposed stream are straightforward from the fact that the auto-covariance function of the superposed stream with \( N \) independent sources is the same as that of single sources [105] (thus the same \( T_n \) as in single sources). And we obtain the distribution of \( S_n \) by \( N \) times convolution of the distribution of the single source. In Table 3.10, cell loss probabilities for the number of sources 5 and 10 are investigated with traffic load \( U = 0.8 \). To clearly show the changes of effect of LRD in multiple-scale buffer sizes, we divide the table into three sections: small (\( K \leq 1,000 \) cells), medium (\( 1,000 < K \leq 10,000 \) cells), and large (\( 10,000 < K \leq 100,000 \) cells) sizes. We use \( h = 100, 1000, \) and 2000 cells for the small, medium, large buffer sizes, respectively. Because of the smoothness due to multiplexing, the effects of hyperbolic correlation (LRD) in the SLCC does not appear clearly in small and medium size region. On the other hand, at large buffer sizes we again find significant difference in cell loss probabilities of the SLCC and the C-DAR(1). However, we defer a definite conclusion for real traffic, since we cannot obtain stable values for the cell loss probability of the empirical trace in less than \( 1.0E - 5 \).

Finally, we verify the accuracy of SL process based modeling by presenting some more numerical results with GOP-smoothed MPEG, and VIC trace A, and B. In Figs. 3.10, we used 64 homogeneous sources for each trace at a traffic load 0.8. All analysis results show queue tail distribution in the accuracy of order. But, the real trace modeling is still challegning, because we need an efficient reduction method
for the distribution for multiplexed sources. In doing applying reduction method, we found that the accuracy of distribution probability at region near the link rate significantly affects the offset of queueing tails. Robust parameter estimation will be critical on the final performance of performance estimation.

Major conclusions that can be drawn from the numerical results in this section are as follows: The SLCC provides accurate and consistent estimation for queueing performance measures, and thus histogram and ACF play a key role in queueing behavior. And the C-DAR(1) model, which does not capture the long-term correlation (LRD) structure of video traffic, also estimates fairly well in light and moderate traffic load and even heavy traffic load with multiplexing video sources.

3.6 Chapter Summary

In this work we investigated the effects of SRD and LRD components in VBR video traffic on queueing performance. We observed that the ACF of an empirical trace is accurately captured by a compound function of the exponential and hyperbolic. To differentiate the effects of the exponential (SRD) and hyperbolic correlation (LRD), we constructed the shifting-level process with compound correlation (SLCC) and presented an efficient and accurate parameter matching algorithm. Especially, the C-DAR(1) model [48] (a SRD video traffic model) is just a kind of SL process with only exponential tail in the SLCC.

We devised a queueing analysis algorithm named the quantization reduction method for the SL/D/1/K queueing system. The application to video traffic showed that the quantization reduction method provides efficient and accurate approximation and the upper and lower bounds of the approximation as well. Especially in the SL process input, the algorithm does not suffer from the computational load
Figure 3.10: Queue tail: 64 multiplexed GOP smoothed MPEG (Top), VIC trace A (Middle), and B (Bottom), utilization of 0.8
because in the SL process the dimension of the one-step transition matrix does not depend on the state size of the input process.

Simulation results showed that the SLCC with the proposed parameter matching algorithm consistently provides accurate prediction of the actual queueing performance. In contrast, the C-DAR(1) model, while showing acceptable prediction under most conditions, underestimates the queue occupancy at large queue sizes under a heavy traffic load. The hyperbolic tail at large lags (LRD) strongly affects the probability in large queue sizes, but only slightly in small buffer sizes. This is why we can find many seemingly contradictory arguments on the importance in the literature.

In this chapter, we did not give a clear expression for dividing the LRD-dominant region and the SRD-dominant region. It should be noted that the region under which a SRD model works successfully depends on the type of video traffic, especially the video coding algorithm. We will consider this in our future work, where real-time parameter estimation of the SL process and on-line admission control based on the SL process will be studied.
Table 3.10: Cell loss probabilities for superposed sources (5, 10) at $U = 0.8$

<table>
<thead>
<tr>
<th>Queue Size (cells)</th>
<th>CLPs for 5 superposed sources</th>
<th>CLPs for 10 superposed sources</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real</td>
<td>SLCC</td>
</tr>
<tr>
<td>200</td>
<td>2.7478E-3</td>
<td>3.4878E-3</td>
</tr>
<tr>
<td>400</td>
<td>7.6870E-4</td>
<td>8.3803E-4</td>
</tr>
<tr>
<td>600</td>
<td>5.2454E-4</td>
<td>5.2767E-4</td>
</tr>
<tr>
<td>800</td>
<td>4.1048E-4</td>
<td>5.1766E-4</td>
</tr>
<tr>
<td>1000</td>
<td>3.7405E-4</td>
<td>5.0800E-4</td>
</tr>
<tr>
<td>4000</td>
<td>2.1745E-4</td>
<td>4.0787E-4</td>
</tr>
<tr>
<td>6000</td>
<td>1.6288E-4</td>
<td>3.6560E-4</td>
</tr>
<tr>
<td>8000</td>
<td>1.1068E-4</td>
<td>3.3398E-4</td>
</tr>
<tr>
<td>10000</td>
<td>9.4143E-5</td>
<td>3.0948E-4</td>
</tr>
<tr>
<td>20000</td>
<td>7.8848E-5</td>
<td>2.3939E-4</td>
</tr>
<tr>
<td>40000</td>
<td>5.6234E-5</td>
<td>1.8287E-4</td>
</tr>
<tr>
<td>60000</td>
<td>0</td>
<td>1.5589E-4</td>
</tr>
<tr>
<td>80000</td>
<td>0</td>
<td>1.3912E-4</td>
</tr>
<tr>
<td>100000</td>
<td>0</td>
<td>1.2734E-4</td>
</tr>
</tbody>
</table>
Chapter 4

The Cutoff Time-Scale: the Relevant Time Scales of Video Traffic for Queueing Behavior

4.1 Introduction

In this chapter, we introduce the cutoff time-scale \(^1\) and develop related queueing theory.

Resource allocation is necessary for a network which guarantees quality of service (QoS). To achieve this QoS guarantee, we must characterize traffic sources so that we can predict the QoS and allocate network resources efficiently. In general, the focus of traffic characterization and modeling lies not in precise modeling of traffic statistics, but rather in capturing the relevant characteristics so as to allow prediction and efficient management.

Variable bit-rate (VBR) compressed video traffic is expected to be a significant component of the traffic mix in integrated service networks. However, the burstiness and delay sensitivity of VBR video poses a severe challenge.

\(^1\)An early version of this work was published in [5]. There we used cutoff interval instead of cutoff time-scale. In this dissertation, we use cutoff time-scale for consistency with dominant time-scale in the next chapter.
Recent research has demonstrated that a compressed video source shows burstiness over multiple time scales [37, 63, 110]. VBR video traffic shows burstiness not only over a period of milliseconds to seconds, corresponding to variations within a scene, but also over a period of tens of seconds to minutes, corresponding to scenes with differing information contents.

Recently, it has been reported that VBR video traffic exhibits long range dependence or asymptotically second-order self-similarity [16, 37, 49], in the sense that the autocorrelation function $\rho(k)(k = 1, 2, \cdots)$ has an infinite sum $\sum_k \rho(k) = \infty$. The relevance of LRD properties with cell loss and delay was studied in [57, 78], and it was argued that there is a critical time scale that has a significant impact on queueing behavior.

A rich set of literature exists on characterizing VBR video traffic with a stochastic process, for example Poisson, Markov modulated, autoregressive, TES, and self-similar (see [95], and references therein; more recently, see [63, 77]). Such stochastic models of sources have the advantage that they may be used to achieve potentially higher network utilization by exploiting the statistical properties of the sources. First- and second-order statistics, i.e., rate distribution and autocorrelation (equivalently, the power spectrum) are considered to be the most important characteristics [26, 44, 95]. However, for network performance estimation, approaches on two extreme conditions are now available. One is based on the bufferless multiplexor model, where the tail distribution of the aggregate rate of multiplexed input is crucial to cell loss probability. The Chernoff bound is one of the most effective methods for call admission control (CAC) under this assumption [29, 95]. The other approach, the effective bandwidth (EB) theory, assumes a very large buffer condition, although it considers the fluctuation of traffic in time scales [95]. However, traffic charac-
4.1 Introduction

terization and CAC for a moderate buffer size is not yet developed. Moreover, it is difficult to implement a policing mechanism based on stochastic characterization and give significant bounds of performance measures in the network.

For these reasons, in this work we consider deterministic traffic characterization and its relationship with network performance. Recently, traffic control based on regulated sources has begun to receive increased attention [12, 13, 14, 18, 19, 20, 27, 41, 72, 86, 87, 96, 95]. Thus, ITU-T and ATM Forum [10, 52] define the rule-based and deterministic traffic contract descriptors: peak cell rate (PCR), cell delay variation tolerance (CDVT), sustainable cell rate (SCR), and maximum burst size (MBS). Further, they define the generic cell rate algorithm (GCRA) for conformance testing of these parameters. An exhaustive survey and comparison of previous studies on guaranteed performance service disciplines with constrained burstiness sources is found in [117]. Especially in [27, 86, 87], traffic sources are regulated according to the leaky bucket constraint. Generalizations of this approach have recently been defined in [13, 14, 19, 20]. By constraining the burstiness of input sources, the network can provide performance guarantees such as buffer backlog and delay bounds. In [58, 113], a set of specific traffic characterization methods for VBR video sources are presented and compared with previous methods.

In this chapter, we concentrate on the time-scale problem in characterization of VBR video traffic: over what range does the traffic correlation have to be considered for estimation of queueing behavior? We first consider the Reich formula and characterize traffic by the piecewise linear arrival envelope function (PLAEF). We then introduce the concept of the cutoff time-scale above which the correlation of input traffic does not affect queue buildup. Since the results are based on deterministic characterization, the cutoff time-scale provides a fundamental bound of the
time scale relevant to the estimation of queue size in the sense that it is free from the possible ambiguity in stochastic approaches for the same issue. The results can be used to determine the meaningful range of correlation for the characterization of VBR video traffic. We quantify the relationship using a set of experiments with compressed video traces by approximating the actual traffic profile with the PLAEF. From mathematical results and simulations, we show that the critical time-scale depends on the traffic load: as the traffic load increases, the range of interval required for estimation of queueing delay also increases.

The study on the time scale relevant to queueing behavior is motivated by the frequently asked question whether LRD in VBR video traffic is crucial for traffic engineering. Our results also offer another insight into the implication of LRD in VBR video traffic.

In addition, we propose an algorithm that approximates the empirical PLAEF to a concave PLAEF, which significantly simplifies the parameter measurement and estimation of cutoff time-scale and delay bound with little estimation loss.

The rest of this chapter is organized as follows. In Section 4.2, we introduce the PLAEF. In Section 4.3, we investigate the calculation of delay bound and cutoff time-scale for a source characterized by the PLAEF. In Section 4.4, we present numerical results of the application of the cutoff time-scale concept, with reference to VBR video traffic sequence. In Section 4.5, we present an algorithm that approximates the empirical PLAEF to a corresponding concave function. In Section 4.6, we conclude our work and suggest further studies.
4.2 The PLA EF: A Deterministic Traffic Characterization

Consider a multiplexer with an output link capacity $C$ and an infinite buffer, loaded by $N$ virtual connections (VC). Let $v_i(s, t)$ denote the number of cell arrivals of $i$-th VC in time $[s, t], i = 1, 2, \cdots, N$ and $v(s, t)$ the overall number of cells arrived at the multiplexer input in the same time interval. By the Reich formula [91]

$$Q(t) = \sup_{s \leq t} \{v(s, t) - C \cdot (t - s)\}, \quad (4.1)$$

where $Q(t)$ denotes the multiplexor buffer occupancy at time $t$. If there exists a bounding function $f(u)$ such that for any $t$ and any $s \leq t$, it can be written $v(s, t) \leq f(t - s)$, then it follows that

$$Q(t) \leq \sup_{s \leq t} \{f(t - s) - C \cdot (t - s)\} = \sup_{u \geq 0} \{f(u) - C \cdot u\}. \quad (4.2)$$

Inequality 4.2 implies that the delay bound is determined by the worst case arrival profile, which may be either for a single traffic $f_i(u)$ or a multiplexed input traffic $f(u)$. The bounding function $f(u)$ can be derived from either constraints imposed on the traffic source at the network ingress or real traffic traces. We resort to the second approach and choose $f(u)$ in the family of piecewise linear functions (i.e., first order splines), hence the name the Piecewise Linear Arrival Envelope Function (PLAEF). The aim in the construction of the PLAEF is using (4.2) to understand the relationship between queueing performance of video traffic multiplexing and the relevant time scale characterization of video traffic source models.

A few deterministic characterizations of VBR video traffic were proposed in [58]. Some of them are equivalent to each other and each model has its advantages in accuracy, policing, easiness for parametrization, or simplicity. The PLAEF directly
models the rate profile curves so that it is very accurate in delay estimation and easy to obtain from the real traces. Furthermore, it has explicit parameters for the time scale of burstiness and correlation of input traffic and thus it is suitable for explaining the effect of the time scale. In the PLAEF, \( f(u) \) is obtained as follows. First, we define \( P \) interval-arrival pairs (IAPs)

\[
\{(t_i, a_i)|t_i < t_{i+1}, a_i < a_{i+1}, (t_0, a_0) = (0, 0), i = 0, 1, 2, \ldots, P - 1\},
\]

(4.3)

where \( a_i \) is the maximum amount of data over any time interval of duration \( t_i \). The PLAEF is defined by the IAPs as the arrival envelope function \( f(u) \) such that

\[
v(s, t) \leq f(t - s) = a_i + \frac{a_{i+1} - a_i}{t_{i+1} - t_i}(t - s - t_i), \quad t_i \leq t - s < t_{i+1}, \quad i = 0, 1, 2, \ldots, P - 1.
\]

(4.4)

For numerical results, we use two differently coded Star Wars traces [37] and 5 MPEG-coded traces [98]. We summarize only the salient characteristics of the traces in Table 4.1.

We call the intraframe coded Star Wars trace the JPEG trace, even though its coding is not compatible with the JPEG standards. The JPEG algorithm also uses only the intraframe redundancy, and thus we expect that the JPEG trace shows characteristics and results similar to a JPEG-coded trace. The other traces are coded according to the MPEG-1 video compression standards with the GOP structure IBBPBBBPPBBBP, where: (i) I denotes an intra-frame coded frame (i.e., without motion compensation); (ii) P denotes a predictive frame, where only the prediction error is coded and prediction is based on adjacent I or P frames; (iii) B denotes Interpolative frames, whose coding is based on interpolation between adjacent I and P frames.

Let \( X(i), i = 1, 2, \ldots \) be the amount of bits generated in the \( i \)-th frame. Because
we have the data in units of bit per frame, we assume that the sources generate the data evenly over the frame interval $T$. For numerical results, we assume all $T = 1/25$ sec. The arrival function for a video trace $v(s,t)$ is given by

$$
v(s,t) = \left( \left\lceil \frac{s}{T} \right\rceil - \frac{s}{T} \right) X(\left\lceil \frac{s}{T} \right\rceil) + \sum_{i=\left\lfloor \frac{T}{T} \right\rfloor +1}^{\left\lfloor \frac{t}{T} \right\rfloor} X(i) + \left( \frac{t}{T} - \left\lfloor \frac{t}{T} \right\rfloor \right) X(\left\lfloor \frac{t}{T} \right\rfloor) \tag{4.5}
$$

where $\left\lceil x \right\rceil$ denotes the ceiling of $x$, i.e., the least integer not less than $x$ and $\left\lfloor x \right\rfloor$ stands for the floor function of $x$, i.e., the largest integer not greater than $x$. The values of the IAPs for the PLAIF are obtained by

$$
t_i = iT, \quad a_i = \max_{s \geq 0} v(s,s+t_i), i = 1, 2, 3, \ldots \tag{4.6}
$$

This is a first order spline approximation of the minimal arrival curve defined in [13], i.e., the function

$$
\gamma(t) = \sup_{s \geq 0} v(s,s+t). \tag{4.7}
$$

In [13] it is shown that $\gamma(t)$ is the least function that upper bounds $v(s,t)$ for any $s$ belonging to the interval $[0,t]$ and for any $t \geq 0$.

The fact that the maximum value in (4.6) is obtained at points of the type $s = jT$ for the video traffic can make the computation easier. If we let $m(i)$ be the index such that the maximum of $v(s,s+t_i)$ is achieved at $s = m(i)T$, then

$$
a_i = \sum_{k=m(i)+1}^{m(i)+i} X(k), i = 1, 2, \ldots P. \tag{4.8}
$$

Figures 4.1 and 4.2 show the PLAIFs for the MPEG trace and the JPEG trace, respectively. The PLAIF for the MPEG trace without any smoothing algorithm shows a periodic pattern due to the GOP structure of its coding algorithm and thus the PLAIF is not a concave function. On the other hand, the PLAIF for

---

\[ \text{Footnote: We have the amount of bits in the unit of the slice for the JPEG trace, but not for the MPEG traces.} \]
the JPEG trace looks like a concave function (in fact, it is not a concave function, either). The concavity of the PLAEF makes the calculation of the delay bound and cutoff time-scale easy. We will consider this in Section 4.5.

4.3 Range of Correlation Relevant to Delay Bound Estimation

We now consider the relationship between the PLAEF and delay bound estimation. We here assume first-come-first-served (FCFS).

Figure 4.3 shows clearly how the PLAEF can be used to assess the worst-case backlog of a multiplexer fed by the corresponding arrival pattern over any given time scale. Let $C_H$ and $C_L$ denote the output link rates ($C_H < C_L$). The subscript $H$ and $L$ mean heavy load and light load, respectively. Although in a practical situation, the link rate is fixed, here we change the link rate $C$ equivalently for simplicity of explanation. In Figure 4.3, we use the MPEG coded Star Wars trace, and one third and a half of the peak rate of the trace for $C_H$, and $C_L$, respectively. This also confirms the intuition that a backlog can be found only up to a critical time scale beyond which no further overload is to be expected. This leads to the following definition of the cutoff scale.

Definition: Let $f(t)$ be an arrival envelope curve of the traffic flow feeding a multiplexer with output capacity $C$; the cutoff time-scale $\tau_c = \tau_c(f, C)$ is given by

$$\tau_c(f, C) = \sup\{u | u \geq 0 \text{ and } f(u) \geq Cu\}. \quad (4.9)$$

The existence of a finite value of $\tau_c$ can be proved as follows. It can be assumed that $f(t)$ is subadditive and $f(0) = 0$. If this were not the case, $f(t)$ could be replaced with another arrival envelope curve, $g(t)$, namely the subadditive closure of $f(t)$, such that $g(t) \leq f(t)$ for any $t \geq 0$ and such that $g(t)$ is subadditive and
4.3 Range of Correlation Relevant to Delay Bound Estimation

Figure 4.1: The PLEAF, mean arrival, peak arrival functions for the MPEG trace

Figure 4.2: The PLAEP, mean arrival, peak arrival functions for the JPEG trace
\[ g(0) = 0 \ [13, 19]. \] In \[18\] it is shown that for the stability of a queue whose input is bounded by \( f(t) \), it must be
\[
\lim_{t \to \infty} \frac{f(t)}{t} = a < C, \tag{4.10}
\]
provided that \( f(t) \) is subadditive and wide-sense increasing. By the definition of limit, it follows that there exists a finite \( t_0 \) such that \( f(t) < Ct \) for any \( t > t_0 \). This implies that the cutoff time-scale \( \tau_c(f,C) \) is finite, as long as the considered multiplexer is stable. In the special case of the PLAEF as defined in Section 4.2, \( f(t) \) is actually wide-sense increasing and subadditive and there exists a finite value for the cutoff time-scale or equivalently a cutoff index \( r \) such that
\[
t_{r-1} < \tau_c(PLAEF,C) \leq t_r. \tag{4.11}
\]
The cutoff index \( r \) can also be defined as the maximum index \( r \geq 1 \) such that
\[
\frac{a_r}{r} \leq C \cdot T < \frac{a_{r-1}}{r-1}. \tag{4.12}
\]

We give a few properties of the cutoff time-scale here:

**Property 4.1:** The queueing delay does not depend on the values of the PLAEF for \( t_i > \tau_c(PLAEF,C) \).

*Proof:* Property 4.1 comes directly from the definition of the cutoff time-scale. \( \square \)

**Property 4.2:** The cutoff time-scale is a decreasing function of \( C \).

*Proof:* Property 4.2 comes from the fact that the PLAEF is a non-decreasing function:
\[
\{t|f(t) \geq CHt\} \supset \{t|f(t) \geq CLt\} \text{ for } CH < CL. \tag{4.13}
\]

\( \square \)

**Property 4.3:** The queueing delay is a decreasing function of \( C \). *Proof:* The proof
is also straightforward from the non-increasing property of the cutoff time-scale with respect to $C$:

$$\sup_{0 \leq t \leq \tau_c(f,C_H)} (f(t) - C_H t) \geq \sup_{0 \leq t \leq \tau_c(f,C_H)} (f(t) - C_L t) \geq \sup_{0 \leq t \leq \tau_c(f,C_L)} (f(t) - C_L t).$$  

(4.14)

Property 4.1 means that the range that is related to the delay of the queue is hard limited by the critical point $\tau_c(f,C)$ and thus a correlation structure longer than this time-scale is of no importance at least in the estimation of queueing performance. Property 4.2 means the range of important correlation decreases as the link utilization increases.

The cutoff time-scale is in essence the largest value of the time scale for which the backlog in a queue receiving an input flow bounded by a function $f(u)$ and emptied with a capacity of $C$ can be non-negative. Then, if one considers such a queue and assumes that it is empty at time $t$, the worst case arrival pattern is such that the queue is empty again no later than $t + \tau_c(f,C)$. Hence, the cutoff time-scale $\tau_c(f,C)$ is an upper bound for the duration of the queue busy period. As such, it also represents the upper bound of delay experienced by any packet fed into the queue, regardless of the serving discipline. Chang [18] noticed such a thing: Here we develop a consistent framework that gives new insight to the remark in [18], by defining the concept of the cutoff time-scale and investigating its properties and applications.

4.4 Application of Cutoff Time-scale to VBR Video Traffic

This section presents numerical results of the application of the cutoff time-scale concept, with reference to VBR video traffic sequences. Figure 4.4 and 4.5 show
Figure 4.3: The illustration of the cutoff time-scale, comparing heavy load and light load condition \((C_H < C_L)\)

cutoff time-scales and link utilizations against the allowed maximum delay bound for a connection, \(D\). The cutoff time-scale and the link utilization are obtained by calculating the minimum link capacity \(C_{\text{min}}\) that guarantees the maximum delay bound \(D\). The minimum link capacity \(C_{\text{min}}\) is given by

\[
C_{\text{min}} = \sup_{u \geq 0} \left\{ \frac{f(u)}{u + D} \right\},
\]

and again the supreme value for the PLAEF is achieved at points of the type \(u = jT, j = 0, 1, 2, \ldots\). The simulation results show that the possible link utilizations are rather low for the deterministic QoS guarantee: utilizations turn out to be 0.2 for the MPEG trace and 0.4 for the JPEG trace for a delay bound of 100 msec. The reason the utilization for the MPEG trace is lower than that for the JPEG trace is that the MPEG is more bursty than the JPEG. However, noting that the link rate is scaled by the input average, the number of MPEG connections admissible for the same link rate is larger than that of the JPEG connections. In addition, the cutoff
time-scale curve shows a step-wise increase for a low delay bound value. This is also
due to the periodic burstiness of the GOP structure. The cutoff time-scale increases
as the traffic intensity increases. Even with deterministic QoS, the cutoff time-scale
is less than 50 frames for traffic loads of 0.5 (the JPEG trace) and 0.3 (the MPEG
trace).

Figure 4.6 shows the cutoff time-scales derived from the PLAEF as a function
of mean load for a multiplexer loaded with: (i) an MPEG-1 coded video trace
representing video clips (MTV2), (ii) the superposition of 20 different MPEG-1
coded video traces with phased GOP, (iii) the superposition of the same 20 MPEG-
1 coded video traces as in (ii) with random phasing. The 20 traces employed here are
available and described in [98]. This figure shows clearly that the cutoff time-scale
beyond which correlations do not matter for queueing performance can be very small
(less than 10 frames, i.e., 400 msec) even for mean loads of 0.6, if the superposition
of independently phased MPEG-coded video streams is considered.

One specific but very important problem closely related with our result is the
frequently asked question of the significance of LRD on traffic control [32, 82, 83].
For example, whether you should incorporate LRD modeling or can use Markovian
model. LRD has been accepted as an inherent characteristic in VBR video traffic,
and it has been reported that the LRD characteristic may impact queueing behavior
significantly [83, 32]. However, several studies have argued that LRD is not crucial
in determining the queueing behavior of VBR video sources [29, 49, 33]. Our result
on cutoff time-scale of video trace explains that the significance of LRD depends on
the input traffic load. More specifically, in the previous chapter we have observed
that autocorrelation function of most available video traces show exponential decay
(implying short-range dependence) up to lags of 30 to 80 frames and hyperbolic
Figure 4.4: The cutoff time-scale variation for the MPEG trace against the maximum delay bound

Figure 4.5: The cutoff time-scale variation for the JPEG trace against the maximum delay bound
Figure 4.6: Cutoff time-scale versus the mean load of the multiplexer, for various input traffic mixes.

decay at more than lags of 30 to 80 frames. Therefore, if cutoff time-scale locates in the SRD region, then LRD affects the queue buildup significantly. Otherwise, the effect of LRD can be ignored.

4.5 Approximation of PLAEF by a Concave Function

In this section, we consider a concave approximation algorithm of the original PLAEF and its advantages in calculation of performance estimation.

Let

\[ m_i = \frac{a_{i+1} - a_i}{t_{i+1} - t_i}, i = 0, 1, 2, \ldots, P - 1. \]  

(4.16)

Then the PLAEF, defined by a set of IAPs \((t_i, a_i), i = 0, 1, 2, \ldots, P\), is a concave function if the sequence of the slope coefficients of the linear pieces \(m_i, i = 0, 1, 2, \ldots, P - 1\), is a non-increasing sequence. If each source has a rate profile that
STEP 1: set $j = 0$, $k(j) = 0$, and $(\bar{t}_0, \bar{a}_0) = (0, 0)$.
STEP 2: $j = j + 1$.
STEP 3: $k(j) = \arg \max_{i > k(j)} (a_i - \bar{a}_{j-1}) / (t_i - \bar{t}_{j-1})$
STEP 4: $(\bar{t}_j, \bar{a}_j) = (t_{k(j)}, a_{k(j)})$.
STEP 5: if $k(j) < P$, go to STEP 2.

Figure 4.7: The Approximation algorithm of an original PLAEF to a concave one
is a concave function, we can simplify the calculation significantly as follows.

**Property 4.4:** If the rate profile function $f(u)$ is a concave function, i) the
largest backlog $Q_{\text{max}}$ is obtained by $f(t_k) - Ct_k = a_k - Ct_k$, where $k$ is the unique
index such that

$$m_{k-1} \geq C \quad \text{and} \quad m_k < C.$$  \hspace{1cm} (4.17)

And ii) the cutoff time-scale $\tau_c(f, C)$ and and the cutoff index $r$ are uniquely deter-
mined by

$$\tau_c(f, C) = \frac{a_{r-1}t_r - a_r t_{r-1}}{C \cdot (t_r - t_{r-1}) - (a_r - a_{r-1})},$$  \hspace{1cm} (4.18)

and

$$a_{r-1} - Ct_{r-1} > 0 \quad \text{and} \quad a_r - Ct_r \leq 0.$$  \hspace{1cm} (4.19)

**Proof:** The location of the maximum backlog is easily obtained if we consider
that $f(u) - Cu$ is also piecewise linear so that it attains its maximum value at the
break point $t_k$ from which the function begins to decrease. The cutoff time-scale is
uniquely obtained by solving $f(t) = Ct$ because of the concavity of $f(u)$. In other
words, the critical time-scale is where $f(t) - Ct$ changes the sign from positive to
negative. Noting $t_{r-1} < \tau_c \leq t_r$ and the linearity in $(t_{r-1}, t_r]$, the equation for the
cutoff time-scale is obtained.

Property 4.4 means that for sources with a concave PLAEF, to obtain the largest
backlog, we need only the arrival information \( a_r \) up to values less than \( Ct_r \) and the link rate \( C \).

When the PLAEF is not a concave function, the backlogs at all time time-scales \( t_i \) have to be compared to obtain the maximum backlog. However, as shown in Section 4.3, the experimental PLAEF is not necessarily a concave function but just a non-decreasing function, especially in the MPEG trace. Therefore, in this section we define an algorithm to derive a concave PLAEF that approximates the empirical PLAEF and examine the difference of the cutoff time-scale estimation and the utilization resulting from the two functions. The algorithm is described in Figure 4.7.

The concavity of the new PLAEF comes from

\[
\frac{\bar{a}_{j+1} - \bar{a}_j}{\bar{t}_{j+1} - \bar{t}_j} < \frac{\bar{a}_{j+1} - \bar{a}_{j-1}}{\bar{t}_{j+1} - \bar{t}_{j-1}} < \frac{\bar{a}_j - \bar{a}_{j-1}}{\bar{t}_j - \bar{t}_{j-1}}. \quad (4.20)
\]

Note also that the new concave PLAEF is an upper bound of the corresponding PLAEF since

\[
\frac{a_i - \bar{a}_{j-1}}{t_i - \bar{t}_{j-1}} \leq \frac{\bar{a}_j - \bar{a}_{j-1}}{\bar{t}_j - \bar{t}_{j-1}} \text{ for } \bar{t}_{j-1} < t_i \leq \bar{t}_j \quad (4.21)
\]

and thus

\[
a_i \leq \bar{a}_{j-1} + \frac{\bar{a}_j - \bar{a}_{j-1}}{\bar{t}_j - \bar{t}_{j-1}} (t_i - \bar{t}_{j-1}) \text{ for } \bar{t}_{j-1} < t_i \leq \bar{t}_j. \quad (4.22)
\]

In Figures 4.8 and 4.9, we compare the results obtained by the concave PLAEF to those by the original PLAEF. For the JPEG trace, the original PLAEF is already nearly a concave function so that the concave PLAEF only slightly differs from the original one. On the other hand, for the MPEG trace the regions of the concave hull are interpolated by a new concave function.

We examine the change in cutoff time-scales (see Figures 4.10 and 4.11) caused by the concave approximation. As we can expect, in the JPEG trace the cutoff
Figure 4.8: The concave PLAEF and original PLAEF for the MPEG trace

Figure 4.9: The concave PLAEF and original PLAEF for the JPEG trace
Table 4.1: Summary of the VBR traces used in the Study

<table>
<thead>
<tr>
<th>Name</th>
<th>Video Source</th>
<th>Length (frames)</th>
<th>Compression Algorithm</th>
<th>peak rate (bits/frame)</th>
<th>mean rate (bits/frame)</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPEG</td>
<td>Star Wars [37]</td>
<td>174,000</td>
<td>intra-frame</td>
<td>627,672</td>
<td>27,791</td>
</tr>
<tr>
<td>MPEG</td>
<td>Star Wars [37]</td>
<td>174,000</td>
<td>MPEG-1</td>
<td>185,267</td>
<td>15,601</td>
</tr>
<tr>
<td>ATP</td>
<td>tennis game [98]</td>
<td>40,000</td>
<td>MPEG-1</td>
<td>190,856</td>
<td>21,890</td>
</tr>
<tr>
<td>MTV2</td>
<td>video clips [98]</td>
<td>40,000</td>
<td>MPEG-1</td>
<td>251,408</td>
<td>19,780</td>
</tr>
<tr>
<td>Talk1</td>
<td>talk show [98]</td>
<td>40,000</td>
<td>MPEG-1</td>
<td>106,768</td>
<td>14,537</td>
</tr>
<tr>
<td>Simpson</td>
<td>cartoon [98]</td>
<td>40,000</td>
<td>MPEG-1</td>
<td>240,376</td>
<td>18,576</td>
</tr>
<tr>
<td>Settop</td>
<td>video-phone [98]</td>
<td>5,000</td>
<td>MPEG-1</td>
<td>46,200</td>
<td>6,031</td>
</tr>
</tbody>
</table>

time-scale for the concave PLAEF is almost the same as the cutoff time-scale for the original PLEAF, especially in the small delay bound region. In the MPEG trace, the concave PLAEF is an upper bound of the original PLAEF, and thus the resulting cutoff time-scale is an upper bound of the original one as well. We observe that the cutoff time-scale does not show a stepwise increase which is due to the GOP structure of the MPEG video coding algorithm.

In Table 4.2, we show the size of the IAP sets \( r \) for the original and concave PLAEFs required to cover the correlation range up to cutoff time-scales for a delay requirement of 10, 50, 100, 500 msec, respectively. We again observe that the concave approximation substantially reduces the number of IAPs for the cutoff time-scale. The IAP set sizes for the MPEG traces are significantly smaller than those of the JPEG traces. Especially, the number of IAPs required for characterizing the relevant time scale region is less than 10 for up to 100 msec but it increases steeply above that point. This is the effect of the GOP structure of MPEG traces. However, the number of 10 even for the concave PLAEF is still large for practical traffic control. Although the focus of this work lies in the charterization problem, in a future work
Figure 4.10: The changes of cutoff time-scales due to the concave approximation for the MPEG trace against the maximum delay bound

Figure 4.11: The changes of cutoff time-scales due to the concave approximation for the JPEG trace against the maximum delay bound
Table 4.2: The Numbers of IAPs $r$ required to cover upto the cutoff time-scale: $r$ for the concave PLAEFs ( $r$ for the original PLAEFs)

<table>
<thead>
<tr>
<th>max. delay (msec)</th>
<th>JPEG</th>
<th>MPEG</th>
<th>MTV2</th>
<th>ATP</th>
<th>Simpsons</th>
<th>Talk1</th>
<th>Settop</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 msec</td>
<td>7(12)</td>
<td>2(2)</td>
<td>2(2)</td>
<td>2(2)</td>
<td>2(2)</td>
<td>2(2)</td>
<td>2(2)</td>
</tr>
<tr>
<td>50 msec</td>
<td>18(23)</td>
<td>5(14)</td>
<td>7(24)</td>
<td>3(11)</td>
<td>3(4)</td>
<td>2(6)</td>
<td>2(4)</td>
</tr>
<tr>
<td>100 msec</td>
<td>18(43)</td>
<td>7(23)</td>
<td>8(38)</td>
<td>3(17)</td>
<td>3(11)</td>
<td>8(86)</td>
<td>4(17)</td>
</tr>
<tr>
<td>500 msec</td>
<td>54(232)</td>
<td>8(81)</td>
<td>16(71)</td>
<td>16(235)</td>
<td>8(82)</td>
<td>19(196)</td>
<td>19(247)</td>
</tr>
</tbody>
</table>

we will examine the effect of matching the original PLEAF to two IAPs, which is equivalent to the present standards, (PCR,CDVT), and (SCR,MBS). We will study this issue in the following chapter.

Finally, to examine the variation and efficiency of the proposed concave algorithm, we show extensive results in Table 4.3. First, we define the gain $G$ as the ratio of peak bandwidth to allocated bandwidth, i.e., $G = \max_i(X(i))/CT$, where $T$ is the video frame time. We find that the JPEG trace has a poorer multiplexing gain than all the MPEG-coded traces for the same delay bounds. Among the MPEG coded traces, Settop and Simpson show a little higher gain than others. Again, we observed that the PLAEF by the concave approximation estimates the cutoff time-scale accurately. For the delay bound of less than 100 msec, the cutoff time-scale is less than 50 frames and very accurate. Even for the delay bound of 500 msec, the error of the cutoff scale is less than 25%. Note that the cutoff index values for the original PLAEF and the concave PLAEF can be much different even though the corresponding cutoff time-scales are almost same.
4.6 Chapter Summary

In traffic theory, rate distribution and time-correlation for input traffic have been considered of first importance for the estimation of network performance. Whether LRD in VBR video traffic is crucial for queueing behavior is also a frequently asked question.

In this chapter, we investigated the relevance between the correlation range and queueing performance, based on a deterministic characterization named PLAEF function. The deterministic characterization, unlike in previous stochastic approaches, can give a fundamental bound of the correlation time-scale relevant to the network delay. We then defined cutoff time-scale above which the correlation does not affect the queue buildup. The cutoff time-scale is the clear bound of the time scale relevant to the estimation of queue size and thus of the meaningful range of correlation for the characterization of VBR video traffic. These theoretical result provides the guideline for the range of time-scales to characterize for static/dynamic Call Admission Control (CAC). The extensive numerical results with MPEG/JPEG traces show the practical range of cutoff time-scales for various video applications.
Table 4.3: Simulation results for various video traces: cutoff time-scales, $t_r$, $\bar{t}_r$ (frames), and link utilization

<table>
<thead>
<tr>
<th>max. delay (msec)</th>
<th>JPEG</th>
<th>MPEG</th>
<th>MTV2</th>
<th>ATP</th>
<th>Simpsons</th>
<th>Talk1</th>
<th>Settop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$t_r$ 7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>G 1.014</td>
<td>1.027</td>
<td>1.032</td>
<td>1.027</td>
<td>1.027</td>
<td>1.032</td>
<td>1.027</td>
</tr>
<tr>
<td>10</td>
<td>$t_r$ 12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 12</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>G 1.052</td>
<td>1.260</td>
<td>1.258</td>
<td>1.252</td>
<td>1.253</td>
<td>1.253</td>
<td>1.260</td>
</tr>
<tr>
<td>20</td>
<td>$t_r$ 16</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 17</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>G 1.093</td>
<td>1.504</td>
<td>1.518</td>
<td>1.504</td>
<td>1.502</td>
<td>1.518</td>
<td>1.502</td>
</tr>
<tr>
<td>30</td>
<td>$t_r$ 21</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 21</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>G 1.093</td>
<td>1.751</td>
<td>1.553</td>
<td>1.770</td>
<td>1.758</td>
<td>1.767</td>
<td>1.776</td>
</tr>
<tr>
<td>40</td>
<td>$t_r$ 22</td>
<td>8</td>
<td>15</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 23</td>
<td>10</td>
<td>19</td>
<td>6</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>G 1.136</td>
<td>2.024</td>
<td>1.938</td>
<td>2.008</td>
<td>2.004</td>
<td>2.016</td>
<td>2.020</td>
</tr>
<tr>
<td>50</td>
<td>$t_r$ 23</td>
<td>14</td>
<td>24</td>
<td>11</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 26</td>
<td>16</td>
<td>28</td>
<td>11</td>
<td>5</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>G 1.166</td>
<td>2.217</td>
<td>2.028</td>
<td>2.273</td>
<td>2.242</td>
<td>2.262</td>
<td>2.252</td>
</tr>
<tr>
<td>60</td>
<td>$t_r$ 24</td>
<td>20</td>
<td>27</td>
<td>13</td>
<td>4</td>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 31</td>
<td>20</td>
<td>32</td>
<td>15</td>
<td>6</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>G 1.190</td>
<td>2.304</td>
<td>2.062</td>
<td>2.513</td>
<td>2.538</td>
<td>2.532</td>
<td>2.551</td>
</tr>
<tr>
<td>70</td>
<td>$t_r$ 25</td>
<td>21</td>
<td>36</td>
<td>15</td>
<td>6</td>
<td>50</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$ 35</td>
<td>21</td>
<td>36</td>
<td>20</td>
<td>8</td>
<td>51</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>G 1.209</td>
<td>2.398</td>
<td>2.092</td>
<td>2.681</td>
<td>2.725</td>
<td>2.747</td>
<td>2.793</td>
</tr>
</tbody>
</table>
Table 4.3: Simulation results for various video traces: cutoff time-scales, $t_r$, $\bar{t}_r$ (frames), and link utilization (continued)

<table>
<thead>
<tr>
<th>max. delay (msec)</th>
<th>JPEG</th>
<th>MPEG</th>
<th>MTV2</th>
<th>ATP</th>
<th>Simpsons</th>
<th>Talk1</th>
<th>Settop</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>$t_r$</td>
<td>35</td>
<td>21</td>
<td>37</td>
<td>15</td>
<td>6</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$</td>
<td>41</td>
<td>22</td>
<td>38</td>
<td>22</td>
<td>9</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>1.229</td>
<td>2.451</td>
<td>2.128</td>
<td>2.747</td>
<td>2.933</td>
<td>2.755</td>
</tr>
<tr>
<td>90</td>
<td>$t_r$</td>
<td>39</td>
<td>22</td>
<td>38</td>
<td>16</td>
<td>9</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$</td>
<td>45</td>
<td>23</td>
<td>39</td>
<td>25</td>
<td>11</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>1.2422</td>
<td>2.506</td>
<td>2.160</td>
<td>2.809</td>
<td>3.086</td>
<td>2.809</td>
</tr>
<tr>
<td>100</td>
<td>$t_r$</td>
<td>43</td>
<td>23</td>
<td>39</td>
<td>17</td>
<td>11</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$</td>
<td>52</td>
<td>24</td>
<td>41</td>
<td>30</td>
<td>14</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>1.256</td>
<td>2.558</td>
<td>2.232</td>
<td>2.874</td>
<td>3.268</td>
<td>2.865</td>
</tr>
<tr>
<td>200</td>
<td>$t_r$</td>
<td>202</td>
<td>27</td>
<td>41</td>
<td>163</td>
<td>67</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$</td>
<td>202</td>
<td>32</td>
<td>51</td>
<td>194</td>
<td>72</td>
<td>116</td>
</tr>
<tr>
<td>300</td>
<td>$t_r$</td>
<td>211</td>
<td>32</td>
<td>47</td>
<td>223</td>
<td>71</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$</td>
<td>212</td>
<td>43</td>
<td>67</td>
<td>226</td>
<td>81</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>1.397</td>
<td>3.279</td>
<td>2.532</td>
<td>3.460</td>
<td>4.587</td>
<td>3.077</td>
</tr>
<tr>
<td>400</td>
<td>$t_r$</td>
<td>220</td>
<td>35</td>
<td>56</td>
<td>229</td>
<td>77</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$</td>
<td>229</td>
<td>56</td>
<td>87</td>
<td>231</td>
<td>92</td>
<td>179</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>1.423</td>
<td>3.584</td>
<td>2.688</td>
<td>3.509</td>
<td>4.785</td>
<td>3.145</td>
</tr>
<tr>
<td>500</td>
<td>$t_r$</td>
<td>232</td>
<td>81</td>
<td>71</td>
<td>235</td>
<td>82</td>
<td>196</td>
</tr>
<tr>
<td></td>
<td>$\bar{t}_r$</td>
<td>242</td>
<td>82</td>
<td>126</td>
<td>237</td>
<td>107</td>
<td>214</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>1.441</td>
<td>3.937</td>
<td>2.865</td>
<td>3.559</td>
<td>4.975</td>
<td>3.226</td>
</tr>
</tbody>
</table>
Chapter 5

Dominant Time-Scale and its Application to Standard-Compatible UPC Selection

5.1 Introduction

In this chapter, we introduce dominant time scale and present an efficient UPC parameter selection method based on the concept \(^1\).

The current standard bodies for broadband integrated service networks, such as ITU-T [52] and ATM Forum [10], use dual leaky buckets to define traffic parameters and conformance test algorithms for real-time VBR service. The IETF (Internet Engineering Task Force) also defines the RSVP (ReSerVation Protocol) for the real time service in future integrated Internet service [102]. The standards basically use the dual leaky bucket model for the traffic specification and usage confirmation, which are defined in terms of PCR (peak cell rate), CDVT(cell delay variation time), SCR(sustainable cell rate), and IBT (intrinsic burst time).

---

\(^1\)A short version of this work appears in [8]. There we used critical time-scale instead of dominant time-scale. In this dissertation, we use dominant time-scale for consistency with cutoff time-scale in the previous chapter.
However, recent research reveals that VBR video traffic shows burstiness in multiple time scales. This implies that, for complete traffic characterization, many traffic parameters are necessary. In particular, Wrege et al. [113] and Ahn et al. [8] studied deterministic multiple time-scale traffic modeling and its effects on traffic control. The simple traffic parameters in standards and the complicated traffic characteristics are conflicting realities, and thus delay VBR video service over high speed integrated networks.

In order to fill this gap, this study suggests a standard-compatible traffic characterization, and investigates its effect on network utilization. Our approach is based on a dominant time-scale, which represents the time scale most important for the queueing performance in a multiplexer. For simplicity in first step of the research, we assume that deterministic call admission control is used [113] and all input has the same traffic characteristics and QoS requirements. More specifically, we use dual leaky buckets \((\rho_1, \sigma_1)\) and \((\rho_1, \sigma_2)\) for traffic parameters.

### 5.2 Dominant Time-Scale in Deterministic Modeling

The dominant time-scale concept has been developed in the probabilistic modeling approach [57, 22] to obtain the time scale dominating the buffer overflow probability at a queue occupancy \(x\) under consideration. The extension of the dominant time scale to the deterministic approaches is simple and straightforward [8]. Figure 5.1 illustrates the dominant time scale \(\tau_d\) in the deterministic approach. Let \(X(i), i = 1, 2, \cdots N\) be the amount of bits generated in the \(i\)-th frame. Because we have the data in units of bit per frame, we assume that the sources generate the data evenly over the frame interval \(T\). For numerical results, we assume all \(T = 1/25\) sec.
First, the $A(t)$ is defined by the maximum arrival:

$$A(t) = \max_{1 \leq k \leq N-t} \sum_{j=k+1}^{k+t} X(j) \quad \text{for } t = 0, 1, 2, \ldots, N$$  \hspace{1cm} (5.1)

We then define the minimum bandwidth $C_{\min}(d)$ for each connection that satisfies the delay bound requirement $d$.

$$C_{\min}(d) = \min\{C > 0 | \max_{t \geq 0} \{A(t) - Ct\} \leq Cd\}$$  \hspace{1cm} (5.2)

The dominant time scale $\tau_d(d)$ is now defined by the time scale where the delay is maximized with the minimum bandwidth $C_{\min}$ for delay requirement $d$.

$$\tau_d(d) = \arg \max_{t \geq 0} \{t | A(t) - C_{\min}(d)t\}$$  \hspace{1cm} (5.3)

The dominant time curves for MPEG coded Star Wars and MTV2 and JPEG-coded Star Wars are plotted in Figure 5.2. The dominant time-scale monotonically increases as the delay bound increase [8], and at small delay bounds, say less than 100 msec, the dominant time-scales for MPEG traces are severely affected by the MPEG coding parameters.

The importance of the concept of the dominant time-scale lies in the fact that the dominant time-scale determines the required bandwidth for meeting delay requirement $d$. This implies that the accuracy of modeling at the dominant time-scale can significantly affect the performance of the multiplexing gain, which is the motivation of the proposed algorithm.

5.3 Leaky Buckets Parameter Selection Algorithm

We now consider usage parameter selection for the dual leaky buckets, $(\rho_1, \sigma_1)$ and $(\rho_2, \sigma_2)$. Wrege et al. [113] proposed an algorithm which reduces the traffic parameters, by approximating the input into a few leaky buckets. Ahn et al. [8]
also proposed a concave approximation for the empirical bound curve. Neither algorithm, however, exploits the concept of the dominant time-scale, and thus no efficient resource utilization is possible. Here we propose an algorithm that maximizes the link utilization on the base of the dominant time-scale. First, we obtain the arrival curve $A(t)$. Then from $A(t)$, we calculate the minimum bandwidth and dominant time-scale $t(d)$. Finally, we construct the dual leaky bucket parameters by the equations.

$$
\rho_1 = \min_{0 \leq t < \tau_d} (A(\tau_d) - A(t))/(\tau_d - t) \quad (5.4)
$$

$$
\rho_2 = \max_{\tau_d < t \leq N} (A(t) - A(\tau_d))/(t - \tau_d) \quad (5.5)
$$

and the cross points with vertical axis is the $\sigma$'s :

$$
\sigma_1 = A(\tau_d) - \rho_1 \tau_d \quad (5.6)
$$
Figure 5.2: The dominant time-scale $\tau_d$ varying the delay bound

$$\sigma_2 = A(\tau_d) - \rho_2 \tau_d$$  (5.7)

Figure 5.3 illustrates an example for the dual leaky buckets for two different delay requirement $d_1$ and $d_2$.

**Property 5.1:** The dual leaky buckets parameters defined in equations (4)-(7) are such that $\rho_1 \geq \rho_2$ and $\sigma_1 \leq \sigma_2$.

**Proof:** From the definition of $\sigma_1$ and $\sigma_2$, it follows that $\rho_1 \geq \rho_2$ implies $\sigma_1 \leq \sigma_2$. So, it is enough to prove the first inequality. From the definition of $t_d(d)$, it follows that $(A(t) - C_{\text{min}}t \leq A(\tau_d) - C_{\text{min}}\tau_d$ for any non negative $t$. Hence, for $t < \tau_d$, we can write $C_{\text{min}}(\tau_d - t) \leq A(\tau_d) - A(t)$ which implies $C_{\text{min}} \leq (A(\tau_d) - A(t))/(\tau_d - t)$ and therefore $C_{\text{min}} \leq \rho_1$. Analogously, the inequality $C_{\text{min}} \geq \rho_2$ can be established. $\square$
It can be also verified that \( A(t) \leq \min(\rho_1 t + \sigma_1, \rho_2 t + \sigma_2) \), i.e. that the dual leaky buckets parameters found by means of (4)-(7) lead to a compliant description of the traffic flow \( A(t) \). First, as a result of Property 5.1, we have that \( \min(\rho_1 t + \sigma_1, \rho_2 t + \sigma_2) \) is equal to \( \rho_1 t + \sigma_1 \) (\( \rho_2 t + \sigma_2 \)) for \( t \leq t_d \) (\( t \geq \tau_d \)). Therefore, it is enough to prove that \( A(t) \leq \rho_1 t + \sigma_1 \) for \( t \leq t_c \) and that \( A(t) \leq \rho_2 t + \sigma_2 \) for \( t \geq t_d \). By the definition of \( \sigma_1 \) and \( \rho_1 \), it follows

\[
\rho_1 t + \sigma_1 = A(t_d) - \rho_1(t_d - t) \geq A(t) \tag{5.8}
\]

An analogous argument applies in case of \( t \geq \tau_d \).

The proposed algorithm based on DTS is optimal in that it provides dual leaky bucket parameters that minimizes required link utilization for given delay bound. This is proved in the following property 5.2.

**Property 5.2:** Let

\[
C^\Psi_{\text{min}}(d) = \inf_{t \geq 0} \left\{ \Psi(t) \right\} \tag{5.9}
\]

Where \( \Psi(t) \) denotes a dual leaky bucket traffic description of a traffic profile \( A(t) \). Then, \( C^\Psi_{\text{min}}(t) \geq C_{\text{min}}(d) \), where \( C_{\text{min}}(d) \) is given by 5.2, for any \( \Psi(t) \) with equality in case \( \Psi(t) \) is given by 5.4 – 5.7.

**Proof:** Since \( A(t) \leq \Psi(t) \) for any \( t \geq 0 \) and

\[
C_{\text{min}}(d) = \inf_{t \geq 0} \left\{ \frac{A(t)}{t + d} \right\}, \tag{5.10}
\]

it follows that \( C^\Psi_{\text{min}} \geq C_{\text{min}}(d) \) for any \( \Psi(t) \). Moreover, by the definition of the dominant time-scale, it is clear that

\[
C_{\text{min}}(d) = \frac{A(t_d)}{\tau_d + d} = \frac{\Psi_{\text{DTS}}(\tau_d)}{\tau_d + d} = C^\Psi_{\text{DTS}}(d), \tag{5.11}
\]
where $\Psi_{DTS}(t)$ denotes the dual leaky buckets traffic description given in 5.4 – 5.7 and the last equality derives from the fact that the dominant time-scale of $\Psi_{DTS}(t)$ is exactly matched to the one of $A(t)$.

Moreover, it can be verified that $C_{\text{min}}^\Psi(d)$ can be computed from the dual leaky buckets parameters and the delay bound according to:

$$C_{\text{min}}^\Psi(d) = \min \left\{ \rho_2, \frac{\sigma_2 \rho_1 - \sigma_1 \rho_2}{\sigma_2 - \sigma_1 + d(\rho_1 - \rho_2)} \right\}.$$

(5.12)

### 5.4 Numerical Results

We conducted comparative simulations of the proposed algorithm and the algorithm used in [113] with the all traces from [38] and [98]. As far as our knowledge, there is no significant results on dual leaky bucket characterization of video traffic except [113] and [8]. Since the proposed algorithm is optimal, performance comparison will show the inefficiency of a blind (without DTS concept) algorithm.

First, as a performance measure, we use the average utilization $\frac{\sum_{i=1}^{N} x_i(t)/N}{C_{\text{min}}^i(d)}$, where the superscript $i$ represents input traffic specification for dual leaky buckets with/without critical time scale. The $C_{\text{min}}^i(d)$ is the minimum link rate that guarantees delay bound $d$ as in (2), but with dual leaky buckets. For the leaky buckets with critical time scale, we used the proposed algorithm. In this case, the same value of $C_i^i(d)$ is obtained as $C_{\text{min}}(d)$ obtained from $A(t)$. For the dual leaky buckets without critical time scale, first two leaky buckets were chosen from multiple leaky buckets [113].

For limit of space, We show only three traces, JPEG-coded Star Wars, MPEG-coded Star Wars, and MPEG-coded MTV2. Other MPEG traces show a little deviation from trace to trace but with the same tendency as shown here. Here we present a property of the dominant time-scale, which verifies the existence and
uniqueness of minimum link allocation.

The multiple leaky bucket modeling is shown in Figure 5.4. To get the parameters of dual leaky bucket, the first two sets of multiple leaky buckets were taken as done in [113]. Note that \( \sigma(i) \) is in decreasing order while the parameter \( \sigma_n \) is in increasing order. The effects of parameters of final two leaky buckets was negligible with a relatively large starting interval \( \tau \), here 200 frames.

Figure 5.5 shows the limit of performance from simply taking the first two leaky bucket parameters of the multiple leaky bucket approximation in [113]. Without the concept of the dominant time-scale, there is no way of focusing on special leaky bucket sets. In this case, we get a close approximation in the small interval ranges, but coarse approximations at the middle and large interval ranges. The effects of a coarse approximation can be found in the simulation results of Figure 5.6. At a small delay requirement the gain of the proposed algorithm is relatively small since
Figure 5.4: The Wrege’s algorithm for multiple leaky bucket approximation

<table>
<thead>
<tr>
<th>Procedure</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> ( A(t) ) and ( \tau )</td>
<td><strong>Output:</strong> ((\sigma_1, \rho_1)) and ((\sigma_2, \rho_2))</td>
</tr>
<tr>
<td>1. Procedure ( \text{DualLB}_\text{Wrege}(A(t), \tau) )</td>
<td></td>
</tr>
<tr>
<td>2. ( i = 0 );</td>
<td></td>
</tr>
<tr>
<td>3. while ( \tau &gt; 0 )</td>
<td></td>
</tr>
<tr>
<td>4. ( i = i + 1 );</td>
<td></td>
</tr>
<tr>
<td>5. ( \sigma(i) = \max_{0 \leq t \leq \tau} \left( \frac{(\tau A(t) - t A(\tau))}{(\tau - t)} \right) );</td>
<td></td>
</tr>
<tr>
<td>6. ( \rho(i) = \frac{(A(\tau) - \sigma(i))}{\tau} );</td>
<td></td>
</tr>
<tr>
<td>7. ( \tau = \min { t \mid \sigma(I) + t \cdot \rho(i) = A(t) } )</td>
<td></td>
</tr>
<tr>
<td>8. end while</td>
<td></td>
</tr>
<tr>
<td>9. ( (\sigma_1, \rho_1) = (\sigma(i), \rho(i)) );</td>
<td></td>
</tr>
<tr>
<td>10. ( (\sigma_2, \rho_2) = (\sigma(i - 1), \rho(i - 1)) );</td>
<td></td>
</tr>
<tr>
<td>11. End Procedure</td>
<td></td>
</tr>
</tbody>
</table>

the dominant time-scale is small, and thus Wrege’s algorithm accurately modeled it; at moderate and large delay requirements, the link utilization gain is significant in the proposed algorithm, since dual leaky buckets without the dominant time-scale concept lack accuracy at those points. This results can be explained in the point of application. For applications with relatively large delay tolerance, e.g. video on demands, time-scale engineering does not give a big benefit, In the other hand, for applications with stringent delay requirement, e.g. video-conference, the burstiness within a few frames dominates the required bandwidth for QoS guarantee.

### 5.5 Chapter Summary

In this paper, we introduced the dominant time-scale in multiple time scale burstiness in VBR video traffic. Based on the dominant time-scale, we proposed an efficient dual leaky buckets parameter selection algorithm which is compatible with
Figure 5.5: Illustration for limit of characterizing input to dual leaky buckets without the dominant time-scale concept.

the current standards. Our proposed algorithm can be directly applied to stored source transmission. In the case of live video transmission, the source may estimate the usage parameters based on the database for the connection type or/and the network operator may estimate on the based of the measured traffic pattern. We leave further investigation of this to future work.
Figure 5.6: Bandwidth utilization gain of the proposed algorithm
Chapter 6

Conclusions

6.1 Contributions of the Dissertation

When we studied video transmission over multimedia networks in the previous chapters, we focused on the multiple time scale burstiness and long-range dependence in VBR video traffic.

We addressed these issues step by step. First, we investigated the origin and detailed statistics of multiple time burstiness and LRD. Then, we proposed a flexible and mathematically tractable traffic model based on the shifting-level process. Through extensive experiments with the SL traffic model and real traces, we fully investigated the effects of these statistics on queueing system. Finally, the cutoff and dominant time-scales was proposed and studied for for call admission control and usage parameter control.

The major contributions of the dissertation are summarized as follows:

- We provided a comprehensive interpretation on the origins and characteristics of multiple time scale burstiness and LRD in VBR video traffic. We recognized video traffic as a hierarchically structured information source, and provided in-depth statistics with various kinds of video traces. This interpretation was
shown to be useful for traffic control. The relationship between the multiple time scale burstiness and LRD is explained in terms of the hyperbolic tail of scene duration distribution.

- We presented a video traffic model based on the shifting-level (SL) process with an accurate parameter matching algorithm for video traffic. The SL process captures all those key statistics of an empirical video trace. Also, we devised a queueing analysis method of SL/D/1/K, where the system size at every embedded point is quantized into a fixed set of values, thus name quantization reduction method. This method is different from previous LRD queueing results in that it provides queueing results over all range not just an asymptotic solution. Further, this method provides not only the approximation but also the bounds of the approximation for the system states and thus guarantees the accuracy of the analysis. Especially, we found that for most available traces their ACF can be accurately modeled by a compound correlation (SLCC): an exponential function in short range and a hyperbolic function in long range. Through intensive and comparative study with DAR(1) model (a well-known SRD model), we found that as far as video traffic loaded queueing system, the rate-distribution and SRD have the first importance for system performance, and the LRD affects significantly only in highly loaded condition.

- We proposed the cutoff time-scale. This cutoff time-scale is defined as the upper bound of the time scale which is required for the estimation of queue size. This concept and related theories give a guideline for traffic modeling research, which is crucial for dynamic or static connection admission control. We also gave a procedure to approximate the empirical PLAEF with a concave
function; this significantly simplifies the calculation in the estimation of the cutoff time-scale and delay bound with little estimation loss. In particular, we showed that the the cutoff time-scale is an increasing function of traffic load and delay requirement. Together with the SLCC (SL compound correlation) model, this theory provide a new insight into the importance of LRD on queueing performance.

- We proposed the concept of dominant time-scale and applied it into the standard-compatible UPC parameters selection. While the standard UPC enforcement is based on the dual leaky bucket, the multiple time-scale burstiness required a large number of leaky buckets for specifying input traffic. The proposed algorithm fills the gap between the standards and reality, by approximating the traffic arrival pattern accurately at the most important time scale, the dominant time scale, for call admission decision. The simulation results with MPEG compressed video trace showed the importance in traffic parameter setup for efficient resource utilization.

### 6.2 Suggestion for Future Work

During the course of this dissertation work, we found several further research issues which are closely related with our study but were not covered properly due to time limit.

In the study of SL process, we focused on a video traffic model. However, the Shifting-level process and its queueing analysis technique is so general that can cover many diverse traffic sources. Modeling Internet traffic such as web, ftp, e-mail by the shifting-level process is a promising area for the further study [103]. Particularly, a specific parameter matching algorithm for this Internet application is required to be
6.2 Suggestion for Future Work

devised. Also, with a multiplexed input of heterogeneous sources, i.e., with different ACF and rate distribution, a SL process cannot match its exact rate-distribution and ACF, even though we cannot get simulation results by generating each sources independently. This difficulty in heterogeneous input is not new but practically meaningful issue.

For cutoff time-scale in Chapter 4 and dominant time-scale in Chapter 5, we took the deterministic QoS guarantee approach. The deterministic approach provided very clear theories and significant benefits. However, we could consider similar concepts and applications with statistical QoS guarantee. This does not mean that our approaches are restricted within deterministic QoS approach. The cutoff and dominant time-scales could be differently defined in statistical QoS approach. Some related studies can be already found in [22, 26, 30, 59, 94].

In addition, the cutoff time-scale can be applied not only to static CAC situation but also to measurement based CAC approaches. We didn’t explicitly investigate this issue but believe that it is another promising research area.

Finally, we would like to mention one general but serious problem in preventive control paradigm, a chicken-and-egg problem. In preventive control, traffic models or UPC descriptors are required for network performance estimation and call admission control. This does not raise any problem in stored video transmission with low interactivity. However, in live video application, it might sound nonsense that we need traffic charcateristics before traffic generation. In such sinarios, we cannot help relaying solely on prediction method: prediction of traffic characteristics, dynamic adjustment of traffic parameters, or/and traffic classification. For example, the effects of prediction error or loss in efficiency could be very practical issues in real network operation.
Bibliography


이력서 (Resume)

Personal
Name
Heejune Ahn (安 照準)
Birth-date and place
26 February 1970 in Seoul
Nationality
Republic of Korea
Military Service
Serving as a Research Officer (EDD: Feb. 2002)
Legal Domicile
15-20 Singongduk-Dong, Mapo-Gu, Seoul, Rep. of Korea
Family status
Unmarried, mother and father and two married sisters

Education
KAIST, PH.D. in Electrical Eng. (Visual Communications)
KAIST, M.S. in Electrical Eng. (Visual Communications)
KAIST, B.S. in Electrical Engineering

Work Experience
Apr. 1996 - present
LGIC (LG Info.& Comm.) Ltd, Research Engineer
Telecommunication Lab, Erlangen-Nuremberg Uni.,
Germany, Visiting Scientist
(on video streaming over mobile networks)
KAIST, Teaching and Researching Assistant

Academic Activity
Research Interests
Visual Communications, Telecommunication Traffic
Control, and Performance Evaluation of Communication
Systems

Membership
A member of KICS, and a student member of IEEE
(since 1996)

Awards
An honor-award in the 5th Samsung Humantech Paper
Contest (March, 1999)
A Gold-award in paper contest hosted by Korean
Ministry of Information and Communication with
LGIC Ltd. (December, 1999)
감사의 글 (Acknowledgments)

밀리 독일에서 지난 5년간의 학위기간을 되돌아 봅니다. 학부의 석사과정까지, 인생의 가장 아름다운 시기를 대전에서 보냈습니. 때로는 기쁨으로 충만했고, 때로는 좌절로 괴로워 했습니다. 그러나, 그 모든 것이 내 삶이고, 내 길이었습니다. 그 길을 가는 동안 그들이 되고 생물이 되어준 많은 분들에게 감사를 드립니다.

우선, 제게 학문의 길을 인도해주신 김재균 교수님께 존경과 감사드립니다. 바쁘신 중에도 논문심사를 해주신 성단근, 채경철, 조동호, 정승, 최봉대 교수님들께 감사드립니다. 특히, 정승 교수님과 최봉대 교수님을 통해 공학적, 수학적 시야를 넓히게 된 것에 감사드립니다. 한국의 멀티미디어 통신을 이끌어 나가고 있는 영상통신 연구실 식구들과 기쁨을 함께하고 싶습니다. 부디 청출어망하고 형보다 아우가 뛰어난 실험실이 되길 바랍니다. 또한, 같이 동거동락 해온 과기대(KIT) 동기들, 이광철, 이용호, 장진원, 여운성, 이상회, 김태석, 김바라에게 고마움을 전합니다. 특히, 이 광철은 여러가지 투정을 목욕히 들어 주었고, 최고의 수학자인 김바라는 항상 저에게 해결책을 제시해 주곤 했습니다. 그리고, 바춰 없이 굴어도 잘 받아주는 선배님들, 강상 سوف, 박성일, 박사, 정인영 박사, 김정곤 박사, 김선용 교수, 설영욱 형, 박강민 형, 또, 언제나 자상하게 예기해 주고 챙겨주시는 민상원 교수. 다들 고마워! 그리고 자주 보자고요. 또한, 박현준, 박성현 교육지도자들과 내가 이뻐하는 동생들, 김원재, 이주웅, 유영민, 박선영, 모두 다 보고싶은 사람들입니다. 또한 LG정보통신의 전현직 연구소장님 및 인사과 직원분들에게도, 입사이후 연구에만 몰두할 수 있도록 경제적, 정신적으로 배려해주신 것에 감사 드립니다.

여기 특별히 감사해야 할 외국인 친구들이 있습니다. I would like to express my special thanks to Ms. Juli Scherer, Prof. Andrea Baiocchi, and Prof. Bernd Girod. Ms. Juli Sherer has taught me the importance of life itself. Prof. Andrea Baiocchi's unsparing comments on my work offered me valuable insights and new directions. Prof. Girod is a real giant both mentally and physically. In addition to his progressive perspectives in professional areas, he gave me the opportunity to discover the western world, Europe. There I met Dr. Frank Hartung and three umlaut friends, Niko Farber, Klaus Stuhlmüller, and Wolfgang Sörgel. I am very grateful to people in Telecommunication Laboratory, Erlangen-Nuremberg University, for their kindness and friendship during my stay in Germany.

마지막으로, 그리고 가장 중요하게, 어머니(차양자) 아버지(안정모)와 우리 광의의 가족(wide-sense-family), 안희욱, 안희경, 김남수, 윤여찬, 김문석, 윤소연, 윤한식과 기쁨을 같이하고 싶습니다. 그들의 배려와 희생이 없이는 불가능한 일이었고, 가능했어야도 의미 없는 일이었을 것입니다.