Traffic Modeling and Performance Analysis for IPTV Systems

by

Fengdan Wan

B.Eng., Southeast University, Jiangsu, China, 2005

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Applied Science

in the Department of Electrical & Computer Engineering

Traffic Modeling and Performance Analysis for IPTV Systems

by

Fengdan Wan

Supervisory Committee

Dr. Lin Cai, Supervisor (Department of Electrical & Computer Engineering)

Dr. T. Aaron Gulliver, Supervisor (Department of Electrical & Computer Engineering)

Dr. Jianping Pan, Outside Member (Department of Computer Science)

Supervisory Committee

Dr. Lin Cai, Supervisor (Department of Electrical & Computer Engineering)

Dr. T. Aaron Gulliver, Supervisor (Department of Electrical & Computer Engineering)

Dr. Jianping Pan, Outside Member (Department of Computer Science)

Abstract

Internet protocol TV (IPTV) is predicted to be the key technology winner in the future. It has, however, stringent quality of service (QoS) requirements. When IPTV traffic shares the network resources with other traffic like data and voice, how to ensure their QoS and efficiently utilize the network resources is a key and challenging issue. In this thesis, Class based queueing (CBQ) is suggested to deploy at the bottleneck router to allow heterogeneous traffic share network resources fairly and efficiently. Then, we propose a two-level Markovian video traffic model and develop a fluid flow based analytical framework to quantify the performance of IPTV systems and derive the admission regions to ensure the QoS of IPTV traffic.

The proposed two-level Markovian traffic model exploits both the temporal and spatial complexity of video traffic. The model can easily be incorporated to network simulators. The fidelity of the proposed video model and the effectiveness of the analytical framework are verified by network simulations driven by real video traces.

Given the traffic model, a fluid flow based analytical framework is developed to study the queueing behaviors of IPTV traffic in wired, single-hop and multi-hop

wireless networks. The analytical results provide insights in how traffic characteristics and network parameters affect the network performance. To ensure the QoS of heterogeneous traffic, admission regions of a wired or wireless bottleneck, with or without CBQ are obtained. The simulation and analytical results both illustrate that time-varying wireless link accommodates much less IPTV connections than the wired link with fixed data rate. The variations of both the incoming traffic arrival rate and the outgoing service rate affect the network performance and admission region. Therefore, it is recommended to deploy proper traffic shaping and resource management schemes in order to support IPTV traffic more efficiently.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACF	Auto-covariance function
AR	Auto-regressive
BP	Beacon period
CAP	Contention access period
CBR	Constant bit rate
CBQ	Class based queueing
CDF	Cumulative distribution function
CTAP	Channel time allocation period
FIFO	First in first out
FSMC	Finite state Markov chain
GoP	Group of Pictures
GPS	Generalized processor sharing
HD	High definition
IPTV	Internet protocol TV
LRD	Long-range dependence
MAC	Medium access control
MES	Minimum error square
NS-2	Network simulator

PLR	Packet loss rate
PNC	Piconet coordinator
QoS	Quality of Service
SNR	Signal to noise ratio
SRD	Short-range dependence
TDMA	Time division multiple access
TES	Transform expand sample
VBR	Variable bit rate
VoD	Video on demand
WPAN	Wireless personal area networks

Acknowledgements

I would like to express my deep and sincere gratitude to my co-supervisor, Dr. Lin Cai, for her understanding, encouraging and personal guidance throughout my study. Working with her for the last two years has been an enjoyable learning experience, though I still have a lot learn from her. I am extremely grateful for everything she has done for me.

I am deeply grateful to my co-supervisor, Dr. T. Aaron Gulliver for his detailed and constructive comments, and for his important support throughout this work. His encouragement and guidance have been of great value in this study. I also learned a lot from him, academically and personally, which I greatly appreciate.

I would like to extend my gratitude to Dr. Jianping Pan who served as my committee member and provided useful insights to my study. I would also like to thank Dr. Alex Thomo for being my external examiner.

My warm thanks are due to my friends, group partners and lab mates for their kind help, suggestions and for giving me a wonderful two years at Victoria.

My special gratitude is to my grandparents, parents, aunts, uncles and cousins for their affection, encouragement, enduring support and confidence in me.

Dedication

To my parents

Chapter 1

Introduction

In this thesis, an analytical framework combined with a proposed video traffic model is developed for IPTV network performance analysis and simulations. It can be used to investigate the characteristics of video traffic that have the greatest impact on network behavior, and determine the admission region of IPTV connections in a bottleneck to ensure their QoS. Thus, the analytical framework is helpful for developing suitable network technologies for IPTV distribution.

1.1 Motivation

IPTV has been predicted to be a major technology winner. Telecommunication service providers are racing to deliver IPTV/video on demand (VoD), voice, and data, the so-called triple-play services. How to efficiently utilize limited network resources to provide triple-play services with guaranteed QoS is a challenging problem.

Due to the heterogeneous characteristics and different QoS requirements of IPTV, data and voice traffic, appropriate resource management schemes should be deployed in the network supporting multiplexed traffic. The transport layer congestion control protocol was first proposed by V. Jacobson [1] to solve congestion of data traffic in the Internet. In recent decades, to meet the stringent QoS requirements for video ser-

vices, various resource management and scheduling schemes have been proposed [2–9]. The purpose of those mechanisms is to protect stringent QoS requirements of video streams from other lower priority traffic classes (flows). Among them, CBQ is a simple link sharing mechanism to support heterogeneous traffic by guaranteeing each traffic class receiving a fair share of the link bandwidth over a relevant time interval [10]. Different from other scheduling policies [10] such as generalized processor sharing (GPS) [7], or using a set-up procedure such as RSVP [4], CBQ does not track the ongoing flow states, and thus it is easy to implement.

Given the link sharing mechanism, and the throughput of wired and wireless links, an immediate question is how many IPTV connections that can be supported simultaneously with guaranteed QoS, or what is the admission region of IPTV traffic sharing the link? With the highly variable data rates and stringent QoS requirements of IPTV traffic, to answer this question, a quantitative analysis of queue behavior with multiplexed video, data and voice over both wireless and wired links is necessary. This not only helps service providers and consumers to choose the best distribution technologies, but also provides important guidelines for planning future networks for IPTV distributions.

In the literature, intensive research has been conducted to analyze the performance of priority queues and scheduling schemes. Separate queues for two traffic classes were used in [11], where a low priority class can consume residual bandwidth only when a high priority class is not able to consume the entire link bandwidth. This can be seen as a special case of CBQ. The queue distribution of two buffers and admission sets for two classes were obtained [11]. Two traffic classes, modeled as Markov fluid sources, were multiplexed by GPS scheduling in [12]. Performance bounds on queue length for each class were numerically validated in the paper. How-

ever, to the best of our knowledge, the general performance of heterogeneous video, data, and voice traffic over a bottleneck with CBQ is still an open issue. This motivates us to investigate it in this thesis. Our analysis demonstrates the effectiveness and efficiency of CBQ in protecting heterogeneous traffic classes sharing the same link.

On the other hand, another difficult issue in quantifying the performance of IPTV is due to new video compression technologies, e.g., MPEG-4 (H.264). Newer video codecs can achieve higher compression rates, but also introduce higher traffic burstiness (peak to average ratio). The fidelity of traffic models plays a crucial role in network performance analysis and simulation. A good traffic model should not be overly complicated and should provide insights into the statistics of video traffic that have the greatest impact on queue behavior. Video frame size modeling has drawn great attention in recent decades. Among the existing models, Markov chain based models are desirable for queue analysis given the well-established fluid-flow analytical framework [13]. Maglaris *et. al* [13] proposed a mini-source based Markov model to describe the variable source rates of different Groups of Pictures (GoPs). It is effective in obtaining the queue distribution when the buffer size is sufficiently large. However, because the mini source model smooths the data rate over the duration of a GoP and neglects the rate variation within a GoP, it cannot reflect the network performance when the buffer size is not large enough to absorb the burstiness within a GoP, as indicated by the results in [14]. In addition, new video compression technologies, e.g, MPEG-4 (H.264) for IPTV/VoD, can achieve higher compression ratio by exploiting the spatial and temporal correlation of video streams. Thus, a good video source model should be able to capture both types of correlations, and reflect the instantaneous income video traffic rate that has great affect on queue behavior.

This motivates us to improve the existing traffic model for IPTV traffic.

Another issue which affects the wireless network performance is the time-varying channel quality of wireless links. Modern wireless communication technologies can adapt the data rate according to channel conditions to maintain a relatively low transmission error rate. Thus, the achievable throughput over a single-hop or a multi-hop wireless path is highly dynamic due to channel variation and rate-adaptiveness. Based on the classic Gilbert-Elliot channel model [15], a finite state Markov chain (FSMC) wireless channel model is adopted for time-varying, location dependent wireless links. The variations on input traffic arrival rate and output service rate will degrade network performance. This motivates us to develop an analytical framework to consider the dynamics of both the input arrival rate and the output service rate simultaneously.

1.2 Contributions

The main contributions of this thesis are three-fold.

First, we develop a two-level Markovian model for IPTV traffic, considering both temporal and spatial correlation, presented by GoP correlation and frame correlation in the same GoP, respectively. The model contains a GoP-level Markov chain and a frame-level Markov chain, to capture both the inter-GoP and intra-GoP correlation, so it can accurately reflect the queue behaviors with both small and large buffer sizes. The statistic properties of instantaneous incoming video traffic that have greatest impacts on queue behavior are estimated in our model, so the proposed traffic model can be used for network performance evaluation of IPTV systems. The model is simple to incorporate in network simulators, and it can be used to obtain closed-form solutions of queue performance. To verify the fidelity of our video model, the queue

performance with synthetic traces generated by the proposed model is compared to the performance with real video traces. Extensive simulations with the Network Simulator (NS-2) demonstrate the feasibility and effectiveness of the proposed traffic model. Thus, it will be an effective tool for performance evaluation of IPTV services via analysis and/or simulations.

Second, to quantify the number of IPTV connections being supported with satisfactory QoS and determine the network performance with very bursty IPTV traffic, a fluid flow based analytical framework is developed in the thesis. Simulation results with NS-2 validate the accuracy of the analysis. The analytical approach could be applied for different network scenarios considering the properties of the video traffic and the accompanying network characteristics. This can be used to reveal the relationship between system parameters, and to provide insights for service providers of design and plan suitable networks carrying IPTV connections.

Third, to ensure the QoS of heterogeneous traffic sharing the same network, CBQ is suggested to be deployed at the bottleneck link. CBQ is easy to implement and effective to protect different traffic types when congestion occurs. The analytical framework is extended to obtain the admission region of heterogeneous traffic with CBQ. The effectiveness of CBQ is validated by the analysis and simulation results.

1.3 Thesis Organization

The reminder of this thesis is organized as follows. In Chapter 2, we provide some background information including the wireless network environment with its Medium Access Control (MAC) protocol, the Markovian wireless channel model, CBQ, and the classic mini-source traffic model for video sources. In Chapter 3, we propose a simple, two-level Markov chain video model for IPTV sources. Synthetic video

traces can be generated and incorporated into network simulations. The analytical framework used to study the network performance is presented in Chapter 4. A Markov chain is derived to model the bottleneck buffer behavior. The framework can be used in various network scenarios, such as traffic over a wired link, a single-hop wireless link, or a multi-hop wireless path, with or without CBQ. The fidelity of our proposed video model and the effectiveness of the analytical framework are verified by the results from extensive NS-2 simulations driven with real video traces, presented in Chapter 5. We present simulation and numerical results for both the proposed model and the mini-source model, showing the accuracy of the proposed one and the effectiveness of the mini-source model to determine the admission region of IPTV connections. The chapter is summarize with the admission region of IPTV connections in various network settings. We conclude the thesis and give the further research topics in Chapter 6.

Chapter 2

Background and Related Work

Given the explosive growth of data rate in optical communication links, the backbone networks have high capacity to accommodate IPTV traffic. Thus, for end-to-end IPTV distribution, the home network is considered a major bottleneck. Given a heterogeneous wired and wireless network architecture for in-home IPTV distribution [16], the admission region of IPTV traffic in the network, i.e., the number of IPTV connections that can be supported with guaranteed QoS, is limited by the number of connections supported by the bottleneck link. Specifically, the end to end packet loss rate (PLR) of the traffic is largely determined by the loss rate in the bottleneck link. Prediction of the traffic performance through a bottleneck link is extremely important for the study of the overall network performance. In order to evaluate the network performance with IPTV traffic, the PLR in a bottleneck link is thoroughly investigated in the thesis. With the existence of triple-play traffic, i.e., data, voice and video, CBQ is suggested to be deployed at the bottleneck link router (e.g., a home gateway) to support different traffic classes with guaranteed QoS. In this chapter, the wireless home network architecture, a FSMC wireless channel model, CBQ and the classic mini-source video traffic model, are introduced.

The traffic model and wireless channel model reveal the statistics of the traffic source and the wireless channel. Compared to bit by bit (physical layer) simulations, statistical models facilitate packet-level network performance evaluation and analysis. In addition, it provides insights into the most significant traffic properties affecting network performance, given the channel conditions.

2.1 Wireless Network Architecture and Channel Model

Since a wireless channel is time-varying and location dependent, the channel conditions make it less suitable than wired channels for carrying high quality IPTV traffic. Thus a wireless link is very likely to be the bottleneck for traffic in the whole network. The performance of IPTV traffic over a wireless link could mainly determine the performance of the traffic through a hybrid wired and wireless networks. In this section, the wireless MAC protocol used in home network and a FSMC wireless channel model are introduced.

2.1.1 802.15.3 MAC protocol

In the wireless domain, although different technologies use different MAC protocols, the basic architectures and approaches are similar. The standard developed for wireless personal area networks (WPANs) is IEEE 802.15.3. We adopt the architecture and MAC protocol specified in the IEEE 802.15.3 standard in our system model.

In an IEEE 802.15.3 WPAN, several devices can autonomously form a piconet, with one device being selected as the piconet coordinator (PNC). The PNC can collect global information about the piconet and allocate radio resources or schedule channel times to all devices in the piconet according to their requests. All devices can communicate in a peer-to-peer fashion using the allocated channel time. Such a semi-ad hoc setting can provide better QoS than a pure ad hoc network. In the

standard, as shown in Figure 2.1, time is slotted in a superframe structure where each superframe consists of a beacon period (BP), a contention access period (CAP) using CSMA/CA as the access protocol, and a contention-free period named channel time allocation period (CTAP) using Time Division Multiple Access (TDMA). All devices can request channel times through contention in the CAP, and use allocated channel times in the CTAP for transmissions. If the traffic must be relayed with multiple hops, the PNC should allocate separate time slots for each hop. With the PNC to coordinate the resource allocation in the wireless network, we do not need to consider hidden or exposed terminal problem, or the link layer contention problem.

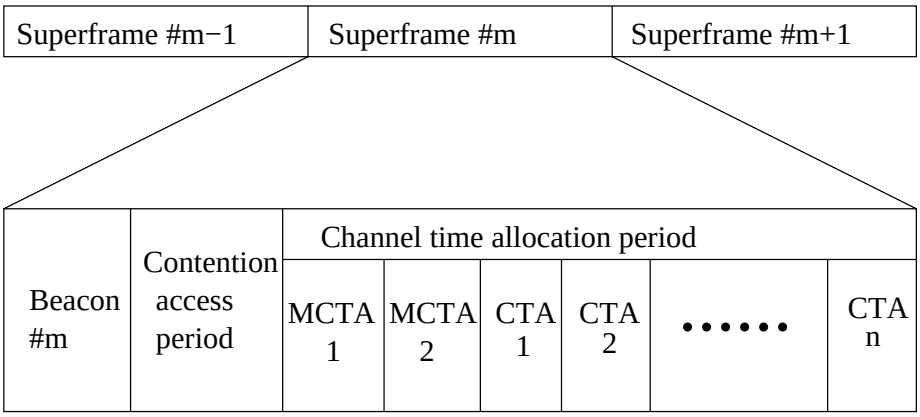

Figure 2.1: Superframe structure defined in the IEEE 802.15.3 MAC protocol.

2.1.2 Wireless channel model (Markov model)

In the wireless domain, advanced wideband wireless technologies can adapt the data rate according to the channel conditions (which are typically time-varying). For instance, the over-the-air data rate of IEEE 802.11n varies from 200 Mbps to 540 Mbps, and that of UWB varies from 53.3 Mbps to 480 Mbps or even higher. With physical

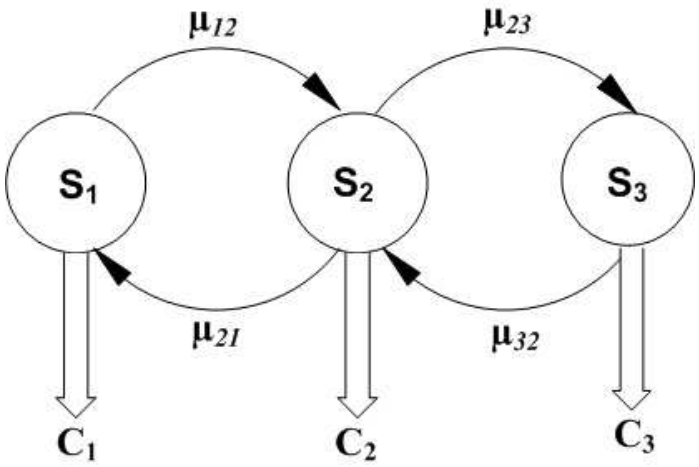

Figure 2.2: Three state wireless channel model.

layer adaptive rate control and link layer retransmissions, from an upper layer protocol's perspective, packet losses due to transmission errors are negligible. However, the link throughput (or data rate) observed by the upper layer protocols is time-varying. We refer to the packet-level wireless channel model as a finite state Markov model [17], because it is recognized to be reasonably accurate in capturing wireless channel variations. With this model, each state corresponds to a different average Signal to Noise Ratio (SNR) and thus a different data rate. The state-transition probabilities are chosen to reflect the time-correlation of the wireless channel. Figure 2.2 depicts a three state Markov chain wireless channel model. The corresponding generating matrix of the chain model can be written as

$$\mathbf{B}_c = \begin{pmatrix} \mu_{11} & \mu_{12} & \mu_{13} \\ \mu_{21} & \mu_{22} & \mu_{23} \\ \mu_{31} & \mu_{32} & \mu_{33} \end{pmatrix},$$

where μ_{ij}, $1 \leq i \leq 3$, $1 \leq j \leq 3$, denotes the transition rate from state i to state j.

2.2 Class-based Queueing Management

With the existence of heterogeneous traffic, to accurately obtain the admission region of triple-play services in a wired or wireless bottleneck path, the traffic characteristics and QoS constraints of supported applications need to be considered. Data traffic is more bursty in nature, and IPTV traffic has very stringent delay and loss requirements. Thus, we propose to deploy a class-based resource management scheme before the bottleneck link. CBQ is a link-sharing approach which enables the gateway to distribute bandwidth on local links in response to local needs [10].

The data packets and video packets are classified into two classes[1]. In the absence of congestion, a general scheduler, first in first out (FIFO) or round-robin scheduler, could be used. When congestion occurs (when one or two classes require more than their allocated bandwidth), CBQ invokes a link-sharing scheduler to rate-limit the over-limit class(es) to their assigned capacity. In this way, CBQ can prevent starvation of data traffic and ensure the QoS requirements of IPTV traffic are met. According to CBQ, the available capacity in the output link for the two classes can be denoted as

$$\begin{aligned} C_v(t) &= \max\{C - \lambda_d(t), (1-p)C\}, \\ C_d(t) &= \max\{C - \lambda_v(t), pC\}, \end{aligned} \tag{2.1}$$

where the arrival rates of IPTV and data traffic are λ_v and λ_d, respectively, and C is the bottleneck link capacity. The instantaneous output rate of each class is bounded

[1]Voice traffic can be represented as another separate class. However, since the volume of voice traffic is much smaller than that of video or data traffic, its impact on the admission region of IPTV connections is negligible, and therefore is ignored in this paper.

by the available bandwidth at any time instant

$$U_v(t) = \begin{cases} \lambda_v(t), & \lambda_v(t) \leq C_v(t) \\ C_v(t), & \lambda_v(t) > C_v(t) \end{cases}$$

$$U_d(t) = \begin{cases} \lambda_d(t), & \lambda_d(t) \leq C_d(t) \\ C_d(t), & \lambda_d(t) > C_d(t) \end{cases}$$

Note that different from the GPS approach (e.g., weighted fair queue), that guarantees the long-term average bandwidth received by different classes, the CBQ scheme adopted here guarantees that each traffic class receives its allocated bandwidth over the relevant time interval [10]. Thus, it can protect IPTV traffic from bursty data traffic. Even if the data traffic is idle for a long period, it cannot occupy more instantaneous bandwidth as compensation (which is not true using GPS), so the instantaneous bandwidth allocated to IPTV traffic is guaranteed. On the other hand, if any traffic class is idle, the other classes can occupy the total link bandwidth to achieve multiplexing gain.

Without tracking the history of on-going flows, CBQ is much simpler to implement than GPS-based schedulers. The different class queues are *virtual* queues for implementation. All packets are multiplexed into one queue with the condition that a certain portion of the data packets is inserted in between video packets. This portion is set according to the link-sharing parameter and the size of the video and data packets. Thus, CBQ can easily be implemented in practice. CBQ is used because of the different characteristics of data and video traffic. In the thesis, data traffic is simply considered as constant bit rate (CBR) traffic or bursty on/off source traffic, while the IPTV traffic characteristics are considered in the video model. After in-

troducing the classic mini-source video model in the next section, we propose a new video model in the next chapter.

2.3 Mini-source Video Model

The statistics of video traffic are determined by several factors, e.g., coding schemes, video content, encoding parameters, and GoP patterns. Extensive research on video modeling has been conducted and a variety of models have been proposed in the literature [13, 18–23]. For video traffic, since the arrival rate changes with respect to frame size and the duration of each frame is relatively long, the bottleneck buffer can quickly reach steady state when different size frames arrive [19]. Thus, the buffer system can be approximated as memoryless, so the steady state of the buffer depends mainly on the current arrival rate.

A classic video model for the source rate is the Markov model based on mini sources [13]. A variable bit rate video source is modeled as 20 multiplexed "on/off" mini sources. Each mini source independently alternates between an "off" state and an "on" state, with A bps generated during the "on" state. The average time in the "off" state and "on" state are $1/\alpha$ and $1/\beta$ seconds, respectively. The number of mini sources in the "on" state follows an underlying Markov chain, as shown in Figure 2.3. The parameter set is obtained by matching the first and second moments of the composite Markov chain model and the measured statistics, i.e., the expectation, variance and auto-covariance function (ACF) of the empirical data [24]

$$\begin{aligned} E(Y) &= MqA, \\ \sigma^2 &= MA^2q(1-q), \\ \alpha+\beta &= K. \end{aligned}$$

The first two equations result from the conservation of the expectation and variance of the Markov chain and the data set, where M is the number of mini sources and $q = \alpha/(\alpha + \beta)$. The last equation comes from matching the ACF of the model to that of the data set. The ACF of a Markov chain is exponential in nature, where $\alpha + \beta$ is the so called the "time constant" of the exponential function. K is chosen by comparing the ACF of the empirical data with the exponential decay function e^K. We find the K that minimizes the total distortion energy, known as the minimum error square (MES) algorithm, defined by

$$\epsilon = \min(\sum_{\tau}(\rho(\tau) - e^K(\tau))^2), \tag{2.2}$$

where $\rho(\tau)$ is the ACF calculated from the data set.

The ACF of a data sequence Y can be defined as

$$\rho(\tau) = \frac{1}{N-\tau}\Sigma_{m=1}^{N-\tau}\frac{(Y_m - E(Y))(Y_{m+\tau} - E(Y))}{\sigma^2}, \tag{2.3}$$

where τ is the time between the two positions and σ^2 denotes the variance. Since the sequence of video frames has the periodicity of the GoP pattern, the ACF of the frame size, according to (2.3) is also periodic, thus it fails to match the aperiodic exponential decay function, and so the mini-source model uses the GoP size as a reference. On the other hand, with the heavy tail property of video traffic, the ACF of the GoP size will not drop exponentially in terms of τ. However, if two GoPs are spaced far apart, their relatively high relation will not affect the buffer occupancy distribution significantly [25]. For instance, three GoPs already span 1.2 seconds of time, which is enough time for the queue states to reach equilibrium. So in (2.2), τ is changing from 0 to 3. Our analytical and simulation results in Chapter 5 suggest

that the cutoff value of $\tau = 3$ is a good trade off between the matching the ACF and preserving the traffic properties affecting queue distribution.

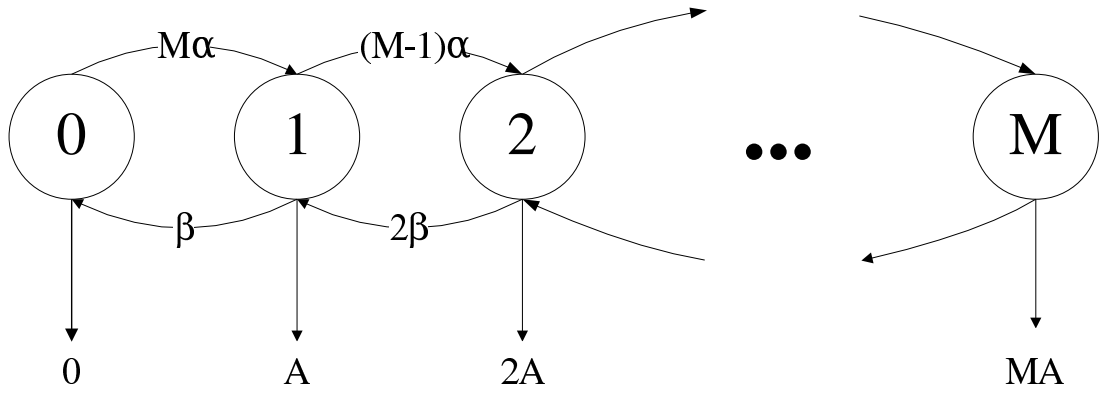

Figure 2.3: The mini-source traffic model.

Due to its high compression and low data rate properties [26], MPEG4 (H.264) video coding technology is considered in the thesis. Even though in the original work [13], the author suggested to use 20 mini sources for each video source, since H.264 has a much higher compression ratio than the source coding considered in [13], we use fewer mini sources to emulate the more bursty video sources using the H.264 codec.

With a Markovian traffic model, the queue distribution is mathematically tractable. However, as a GoP-level model, the mini-source model makes the time scale coarse and so it is not suitable for high speed networks [18]. In [14], simulation results showed that the mini source model underestimates the packet loss rate (PLR) of video sources over a bottleneck link with a small buffer size. Since most video codecs for IPTV use the GoP structure, and intra-GoP correlation plays an important role on the queueing behavior, the original mini source model is not directly applicable for studying IPTV systems.

Newer video models combine a Markov process with other stochastic processes.

For example, Dawood *et al.* [20] grouped a virtual video-clip into a number of shots (states), according to the video content. Then, they specified the shot duration using the gamma distribution, and an auto-regressive (AR) model was used for each shot. In [21], the gamma distribution was chosen for the state-dependent arrival rate, while Transform-Expand-Sample (TES) [22] was proposed to model the duration and bit rate process in each state. Krunz *et al.* [23] used the M/G/∞ process for the model by selecting an appropriate G to capture the marginal distribution of a video sequence. These models mainly focus on representing both short-range dependence (SRD) and long-range dependence (LRD) of a variable bit rate (VBR) video source or accurately matching the marginal distribution of the frame size. However, the partition of states and models used in each state is very complicated. Thus, they are not easy to apply in the mathematical analysis of queue behavior. Other approaches include the wavelet based video model by Dai *et al.* [18], in which the I frame size is modulated by wavelets and synthetic P/B frame sizes are determined using linear functions. This model is also not very suitable for queue behavior analysis because of its high computational complexity.

Since the aim of the video traffic models considered here is to facilitate network performance analysis and simulations, it is desirable to obtain a simple traffic model that can accurately estimate the statistics of the video traffic that have the greatest impact on queue behavior. Different from previous work, our model, proposed in the next chapter, partitions the video frames into different states with a simpler process, and both the inter-GoP and intra-GoP correlations are exploited using a two-level Markov chain. This model is general and can be used for different video sources encoded with a GoP structure. The model can easily be incorporated into network simulators such as NS-2.

Chapter 3

A Simple, Two-Level Markovian Traffic Model

In this chapter, a simple two-level Markovian traffic model is proposed, which considers the temporal and spatial correlations of video, and the inter-GoP and intra-GoP correlations.

For a typical MPEG4 (H.264) video codec [26], a continuous video stream is sampled into a sequence of frames as the input of the encoder. After encoding, frames are emitted periodically comprising GoPs. Each GoP contains an I frame and a number of P and B frames. For example, with an MPEG codec, the generic GoP is $I\ B_1\ B_2\ P_1\ B_3\ B_4\ P_2\ B_5\ B_6\ P_3\ B_7\ B_8$. The first frame in each GoP is an I frame, which is intra-coded without reference to any other frames. The following P frames are both intra-coded and inter-coded with respect to the previous P or I frame. The remaining B frames are also intra-coded and inter-coded, and they use both the previous and subsequent P or I frames as references. Since only I frames are encoded without exploiting temporal redundancy, typically an I frame has the largest frame size in each GoP.

3.1 GoP-level Markov Chain

The main task of video compression is to remove spatial and temporal redundancy within each and between consecutive frames to save bandwidth. Hence frame size depends on the texture and motion complexity of the video content, or its spatial and temporal domain correlations. In the spatial and temporal domains, we categorize the video into a number of levels, S and T, respectively. Thus, we can use $S \times T$ states to represent the correlations in both domains. Since the duration of a GoP is less than half a second, we assume that the spatial and temporal correlations of the video source in each GoP remains at the same level, or in the same state. Therefore, we can build a GoP-level discrete time Markov chain, with each state representing the temporal and spatial correlations of the video within that GoP.

The more states (larger S and T), the more accurate the model, but at the cost of computational complexity in generating and applying the model. Experimental results show that choosing $S = 3$ and $T = 3$ provides a good trade-off between model accuracy and complexity. The three levels are denoted as low complexity (L), medium complexity (M) and high complexity (H) states.

The boundaries between states should be set appropriately, so that we can use a limited number of states to accurately capture the video statistics. According to the experimental results with a few video streams, we note that evenly dividing the frames into each state gave reasonably good results. That is,

$$Pr(S_i) = 1/N,$$

where S_i denotes one of the states, and $N = S \times T$ is the total number of states

In the spatial domain, since only the I frames are independently intra-coded, the

I frame size is used to determine the texture complexity of the entire GoP. In the temporal domain, the ratio of size of the first P frame to the I frame size in the same GoP is used to indicate the temporal correlation, as explained below. The P_1 frame is both intra- and inter-coded with sole reference to the I frame. The frame size of P_1 is determined by

$$P_1 = P_{1t} \times \Delta\varphi,$$

where $\Delta\varphi$ denotes the motion vectors from frame I to frame P_1 in the same GoP, and P_{1t} is the texture information contained in frame P_1, which is approximately the same as the texture information in the I frame of the same GoP. Hence, the ratio between the first P frame size and the I frame size in the same GoP, ϕ, indicates the correlation in the temporal domain

$$\phi = P_1/I = \Delta\varphi.$$

The larger $\Delta\varphi$, the higher the video motion speed, and the lower the correlation in the temporal domain. Thus, the ratio of P_1 to the I frame size in the same GoP can be used to determine the state in the temporal domain.

Combining the three states in the spatial domain with the three states in the temporal domain gives nine states for each GoP, as shown in Figure 3.1. For example, state LM denotes a GoP with low correlation in the spatial domain but medium correlation in the temporal domain. Note that every pair of states is connected. The transition probability between a pair of states is jointly determined by the temporal and spatial transition probabilities. We assume that the temporal and spatial processes are independent. Then the transition probability from state LM to state

MM, for example, is

$$Pr\{LM \rightarrow MM\} = Pr_s\{L \rightarrow M\} \times Pr_t\{M \rightarrow M\},$$

where $Pr_s\{L \rightarrow M\}$ and $Pr_t\{M \rightarrow M\}$ denote the spatial transition probability from the L ("Low") state to the M ("Medium") state, and the temporal transition probability from the M ("Medium") state to the M ("Medium") state, respectively. Each Pr_t and Pr_s is obtained by counting the number of transitions between two assigned states in the trace.

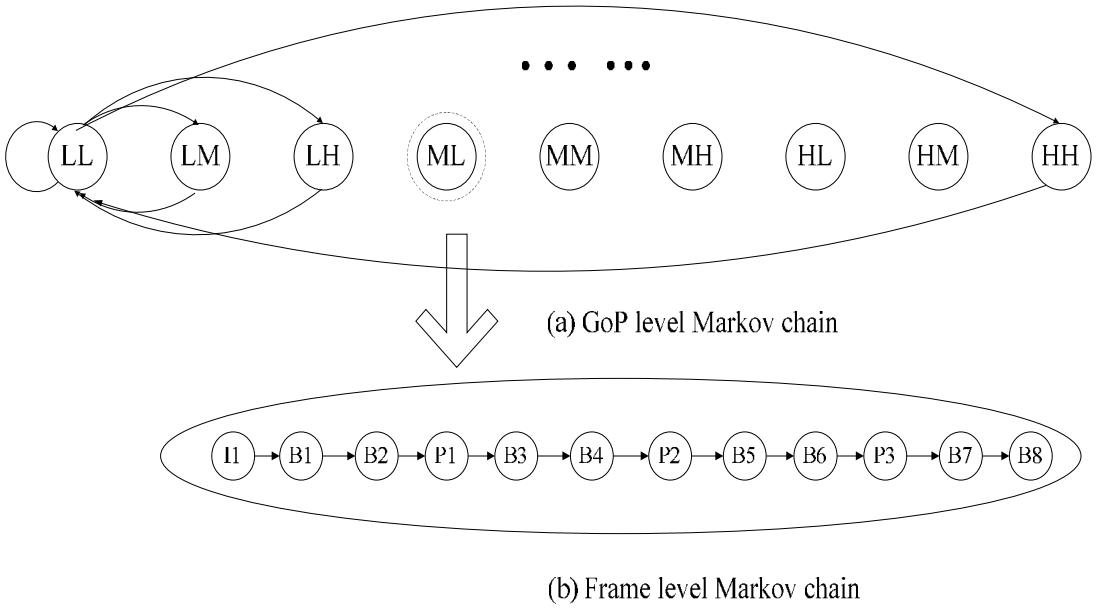

Figure 3.1: A two-level Markovian traffic model.

3.2 Frame-level Markov Chain

Since the GoP-level Markov chain cannot capture the burstiness of the traffic arrival rate within the GoP, we extend the above model to consider different types of frames inside each GoP, and build a discrete time Markov model considering not only the

inter-GoP correlation, but also the intra-GoP correlation.

The time step for the frame-level Markov model equals the duration of a video frame. Each state in the GoP-level model corresponds to a 12-step Markov chain at the frame-level, as shown in Figure 3.1. This represents the 12 frames in the GoP. The state transition probabilities within the GoP are deterministic. Each state corresponds to a different frame type with a different traffic arrival rate. With a limited number of states, a large number of frames with a wide range of frame sizes will belong to the same state. Thus, the critical issue is how to associate the frame sizes with the states.

The size of an I frame is determined by the spatial domain correlation only. Therefore, the I frame sizes in states XL, XM, and XH are the same for $X = L, M \text{and} H$ in the spatial domain. A simple method to determine the frame size of each I frame state is to average the size of all I frames belonging to that state

$$\overline{I^X} = \sum_{I_j \in \{\text{state } XL, XM, XH\}} I_j / N^X, \tag{3.1}$$

where N^X is the total number of I frames in states XL, XM, and XH, and I_j is the I frame size in the j-th GoP. $\overline{I^X}$ in (3.1) is the average I frame size in states XL, XM, and XH, which can be used to represent the I frame size in the corresponding states.

The B and P frame sizes, however, are determined by the combined spatial and temporal state. As discussed before, the ratio of P_1/I indicates the temporal correlation of each GoP. Thus the average of P_1/I is used in the temporal domain state as

$$\alpha_{P_1}^K = \overline{I^K} / \overline{P_1^K},$$

where $\overline{I^K}$ and $\overline{P_1^K}$ are calculated similar to (3.1), but in the temporal domain instead. That is

$$\overline{I^K} = \sum_{I_i \in \{\text{state } LK, MK, HK\}} I_i / N^K, \tag{3.2}$$

where N^K is the total number of I frames in states LK, MK, and HK, and I_i is the I frame size in the i-th GoP. $\overline{I^K}$ in (3.2) is the average I frame size in states LK, MK, and HK. Finally the joint state decides the P_1 frame size as

$$\overline{P_1^{XK}} = \alpha_{P_1}^K \overline{I^X}. \tag{3.3}$$

For the remaining B and P frames, we first examine the intra-GoP correlation. The correlation coefficient of the two sequences i and j is

$$R(i,j) = \frac{CoV(i,j)}{\sqrt{CoV(i,i)CoV(j,j)}},$$

where CoV denotes covariance. The R value between the B/P frames and P_1 in the same GoP are examined. The value of R is less dependent on video content, compression scheme, etc. Typically, there is a high correlation between frames in the same GoP, e.g., the video trace mainly investigated in the paper, "Mars", has the lowest value of 0.75 [27]. Therefore other P/B frame sizes are generated based on (3.3) as well

$$\overline{F_T^{XK}} = \alpha_T^K \overline{I^X}, \text{for } T \in \{B_1, B_2, ..., B_8, P_2, P_3\}. \tag{3.4}$$

where $\overline{F_T^K}$ denotes the average size for each type of fames, and it is calculated similar to (3.1), and K indicates the three states in the temporal domain. Different from the approach in [18] and [28] where the average P and B frame sizes in the GoP are used to represent all P/B frames in the same GoP, we differentiate the sizes of P

and B frames in different locations. This is because HD video has much larger frame sizes than the video data investigated previously. Even though the values of α_T^K for different types of P/B frames may be similar, after multiplication in (3.4) with a large variety of I frame sizes, the discrepancy will be significant.

In summary, the proposed model development has two steps: first, categorize the GoPs into states in the temporal and spatial domains, build a GoP-level Markov chain, and determine the state transition probability matrix from the video trace; second, extend each GoP state to a 12-step Markov chain in the frame-level, and determine the frame sizes of each state based on the average value and linear relationships of the P/B and I frame sizes in the same state.

Remarks: The original mini source model, a GoP-level model, actually resulted in a coarser time scale. The time granularity of the mini source model is equal to the duration of one GoP, e.g., 0.4 s for a 30 frames per second video trace, which is a relatively long time for high speed networks. This model smooths out the video traffic rate over the entire GoP duration, ignoring arrival rate variations within a GoP. For example, the higher I frame data rate is averaged by the low rate of the B and P frames in the same GoP. Thus, the mini source model cannot capture the high loss rate of video packets when the buffer is not large enough to accommodate all packets from a large size I frame. Our model exploits both inter-GoP and intra-GoP correlations, without significantly increasing the complexity of the model.

Another issue worth mentioning is the LRD property of video traffic. The challenge for a video model is to capture both the SRD and LRD properties of VBR video [29]. With just a constant rate modulated Markov process, our model is Markovian in nature with an autocorrelation that drops exponentially. Thus, it cannot capture the LRD of a video trace. However, previous study showed that LRD is

not a crucial property in determining the buffer behavior of VBR video sources [25]. The strong frame-to-frame correlation is more important, while the long-term form of the autocorrelation function, whether it is exponential over several frames or has some periodic component, has a negligible effect on the buffer statistics [19]. Our simulation results with real video traces also demonstrate that the buffer behavior of one or several multiplexed video sources can be accurately determined by using our model.

3.3 Summary

In this chapter, we proposed a simple, two-level Markovian traffic model, considering the spatial and temporal complexity of video traffic. The proposed traffic model considers both GoP level and frame level correlations. The merit of the proposed model, compared with some existing video models, is its mathematical tractability for network performance evaluation and analysis. The statistic properties of instantaneous incoming video traffic that have greatest impacts on queue behavior are captured in our model.

Given the transition matrix of video frames generated by the proposed model, the queue distribution could be derived according to the methodology given in the next chapter. Therefore, given the buffer size of the bottleneck link, we can obtain the maximum number of IPTV connections that can be supported with guaranteed QoS (i.e. the admission region of IPTV systems). In addition, by simply using average modulation rate in each state, the proposed model can be used to calculate effective capacity which is defined and discussed in the next chapter.

Chapter 4

Analytical Framework

To quantify the number of IPTV connections being supported with satisfactory QoS and determine the network performance with very bursty IPTV traffic, an analytical framework based on the fluid flow approach is developed in this chapter. The approach could be applied specifically for different network scenarios, considering the properties of the video traffic and the accompanying network characteristics. Given the video models and the wireless channel model, CBQ management for heterogeneous traffic, the queue distribution is obtained. For the analysis, a simple FIFO queue management scheme is considered, as shown in Figure 4.1. N sources, $S_1, \ldots, S_n$ are multiplexed in a FIFO queue before the bottleneck link C.

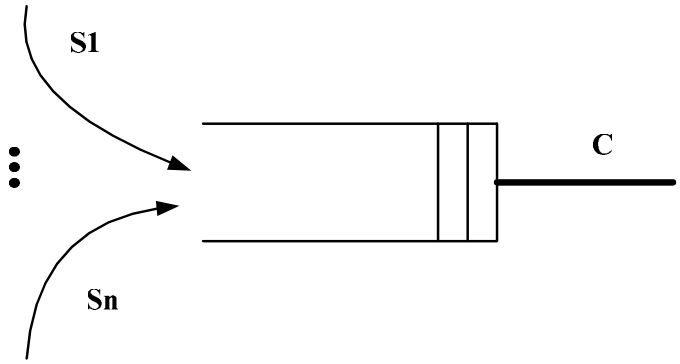

Figure 4.1: The multiplexed queue system.

4.1 Performance Analysis with The Fluid Flow Approach

In this section, the fluid flow based analytical framework is developed. Given the queueing system, i.e., the generating matrix of the underlying Markov chain, arrival rate and service rate at each state, the queue distribution can be obtained by solving a uniform derivative equation. The loss rate of traffic with a given queue length can then be determined. In other words, the maximum buffer size can be used to support IPTV connections with guaranteed QoS. By quantifying queue performance analytically, the admission region of IPTV traffic is obtained. It can help service providers improve the design and deployment of networks to support IPTV traffic and choose a proper resource (buffer and/or bandwidth) allocation scheme. In addition, it provides important insights into which system parameters and/or traffic characteristics affecting the admission region of networks.

Theorem 1 *Let $F_i(x)$ denote the probability that the queue length is less than x, given that the system is in state i, $1 \leq i \leq N$. $\mathbf{F}(x)$ is the row vector*

$$[F_0(x)\ F_1(x)\ ...\ F_N(x)],$$

with the total number of states, N. The equilibrium queue length distribution at the bottleneck link is subject to

$$\frac{d\mathbf{F}(x)}{dx}\mathbf{D} = \mathbf{F}(x)\mathbf{B}, \tag{4.1}$$

where $\mathbf{D}$ is an $(N+1) \times (N+1)$ diagonal matrix

$$\mathbf{D} = \mathrm{diag}\{\mathrm{R}_1 - \mathrm{C}_1, ..., \mathrm{R}_\mathrm{N} - \mathrm{C}_\mathrm{N}\},$$

where R_i and C_i denote the arrival rate and the service rate at the i-th state, respectively. The generating matrix for the underlying (continuous) Markov chain of the system is represented by

$$\mathbf{B} = \begin{pmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1N} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{N1} & \mu_{N2} & \dots & \mu_{NN} \end{pmatrix}, \tag{4.2}$$

where μ_{kj}, $1 \leq k, j \leq N$, is the state transition rate from state k to state j, and $\mu_{kk} = -\sum_{j=1 j \neq k}^{N} \mu_{kj}$.

Proof Let $F_i(t, x)$ be the probability that the queue length at the bottleneck link not exceed x at time t, when the system is in state i. The probability of a state transition from state k to state j during a small time interval Δt is $\mu_{kj}\Delta t$. Then, at time instant $t + \Delta t$

$$F_i(t + \Delta t, x) = \sum_{k=1, k \neq i}^{N} \mu_{ki} \Delta t F_k(t, x) + (1 - \sum_{i=1, i \neq j}^{N} \mu_{ij}) \Delta t F_i(t, x - \Delta x)$$

$$+ \text{ higher order terms of } \Delta t$$

where $\Delta x = (R_i - C_i)\Delta t$ is the net gain of the buffer in the small period t to $t + \Delta t$. The high order terms of Δt is ignored as $\Delta t \to 0$.

Add $-F_i(t,x)$ to both sides of the above equation and divide by Δt to obtain

$$\begin{aligned}\frac{F_i(t+\Delta t,x)-F_i(t,x)}{\Delta t} &= \sum_{k=1,k\neq i}^{N}\mu_{ki}F_k(t,x)-\sum_{i=1,i\neq j}^{N}\mu_{ij}F_i(t,x-\Delta x)\\ &\quad +\frac{F_i(t,x-\Delta x)-F_i(t,x)}{\frac{\Delta x}{R_i-C_i}}.\end{aligned}$$

Now let $\Delta t\to 0$ and $\Delta x\to 0$ giving

$$\frac{\partial F_i(t,x)}{\partial t}=\sum_{k=1,k\neq i}^{N}\mu_{ki}F_k(t,x)-\sum_{i=1,i\neq j}^{N}\mu_{ij}F_i(t,x)+(R_i-C_i)\frac{\partial F_i(t,x)}{\partial x}.$$

The queue is in equilibrium, i.e., $\frac{\partial F_i(t,x)}{\partial t}=0$ and $F_i(t,x)=F_i(x)$. We finally have

$$(R_i-C_i)\frac{dF_i(x)}{dx}=\sum_{k=1,k\neq i}^{N}\mu_{ki}F_k(x)-\sum_{i=1,i\neq j}^{N}\mu_{ij}F_i(x) \tag{4.3}$$

Equation (4.3) is a set of derivative equations for $F_i(x), 1\le i\le N$. Organize $F_i(x)$ as a row vector $\mathbf{F}(\mathbf{x})$ and the coefficients into the matrices

$$\mathbf{D}=\mathrm{diag}\{\mathrm{R_i}-\mathrm{C_i}\},1\le \mathrm{i}\le \mathrm{N}$$

and

$$\mathbf{B}=\begin{pmatrix}\mu_{11} & \mu_{12} & \dots & \mu_{1N}\\ \mu_{21} & \mu_{22} & \dots & \mu_{2N}\\ \vdots & \vdots & \vdots & \vdots\\ \mu_{N1} & \mu_{N2} & \dots & \mu_{NN}\end{pmatrix}.$$

Then, (4.1) is readily obtained. ■

According to [13], the solution of (4.1), i.e., the cumulative distribution function (CDF) of the queue length is

$$F(x) = 1 + \sum_{i:Re[z_i<0]} a_i \sum_{j=1}^{M} \phi_{ij} \exp(z_i x),$$

where z_i and $\mathbf{\Phi}_i$ are the negative left eigenvalues and the corresponding eigenvectors of the matrix $\boldsymbol{BD}^{-1}$

$$z_i \mathbf{\Phi}_i \mathbf{D} = \mathbf{\Phi}_i \boldsymbol{B}. \tag{4.4}$$

The coefficients a_i can be obtained from the boundary conditions, i.e., $F_j(0) = 0$ for $j > C/A$. The survivor function $G(x)$, which represents the probability of buffer overflow, is the complementary distribution of $F(x)$

$$G(x) = 1 - F(x) = - \sum_{i:Re[z_i<0]} a_i \sum_{j=1}^{M} \phi_{ij} \exp(z_i x). \tag{4.5}$$

In summary, given $\mathbf{B}$, the generating matrix of the underlying Markov process representing the variation of effective bandwidth, and the rate matrix, $\mathbf{D}$, the queue length distribution can be obtained. Then by using (4.5), the probability of buffer overflow can be predicted. In the following section, we explain the approach to determine $\mathbf{B}$ and $\mathbf{D}$ for each system (network configuration), from the simplest case with just a wired link to the case of a wireless link and heterogeneous traffic classes with CBQ management.

4.2 Case Study I: Wired Backbone Network

We assume the wired link has constant capacity, thus the underlying Markov chain for the queue system is just the chain modeling the video sources. The matrix $\mathbf{B}$ is the

generation matrix of the video sources. Based on the two video models introduced in Chapters 2 and 3, we have $\mathbf{B}_1$ for mini-source model and $\mathbf{B}_2$ for the proposed model. According to Figure 2.3

$$\mathbf{B}_1 = \begin{pmatrix} -M\alpha & M\alpha & 0 & \dots \\ \beta & -(M-1)\alpha - \beta & (M-1)\alpha & \dots \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & -M\beta \end{pmatrix},$$

and

$$\mathbf{D}_1 = \text{diag}\{0 - \text{C}, \text{A} - \text{C}, ..., \text{M} \times \text{A} - \text{C}\},$$

where M is the total number of mini sources. In this thesis, M is set to 8 for every IPTV stream.

However, based on Figure 3.1, $\mathbf{B}_2$ represents the transition between individual frames. Let $i = 1$ to 9 denote the states "LL", "LM", "LH", "ML", "MM", "MH", "HL", "HM" and "HH" in the model, respectively. Thus, if the states in $\mathbf{B}_2$ are organized as $I^1, \cdots, I^9, P_1^1, \cdots, P_1^9, \cdot, P_3^9, \cdots B_1^1, \cdots, B_1^9, \cdots, B_8^9$, $\mathbf{B}_2$ has the following structure

$$\mathbf{B}_2 = \begin{pmatrix} \mathbf{O} & \mathbf{I} \\ \mathbf{BI} & \mathbf{O} \end{pmatrix},$$

where $\mathbf{I}$ is a 99×99 identity matrix which indicates deterministic transitions within each GoP state. $\mathbf{B}_2$ represents the expanded chain in Figure 3.1, and $\mathbf{BI}$ is a 9×9 GoP-level matrix which is the transition matrix from the last frame (a B_8 frame) in the previous GoP to the I frame of the next GoP. Thus $\mathbf{B}_2$ has the dimension of 108

for every video source.

$$\mathbf{D}_2 = \text{diag}\{\text{R}_1 - \text{C}, \text{R}_2 - \text{C}, ..., \text{R}_{108} - \text{C}\},$$

where R_i is the data rate generated in state i, which equals the frame size of the state over the frame duration.

Given **B** and **D**, using the fluid flow approach, the queue distribution can be obtained. Given the QoS requirements of IPTV traffic, including the loss rate and delay bound, we can limit the number of connections and choose an appropriate buffer size. For instance, given the delay budget in a home network, we can determine the required queue length and thus the buffer size. From (4.5), we can determine the maximum number of connections that can be supported with a guaranteed loss rate due to buffer overflow.

4.3 Case Study II: Wireless Network

To quantify the admission region of IPTV traffic with the existence of data traffic, identify the bottleneck, and study the traffic performance in the heterogeneous wired and wireless mesh network, the fluid model is extended to consider a time-varying single-hop wireless link with a variable data rate. This is later extended to multi-hop wireless paths. We study the queue behavior and admission region for homogeneous IPTV traffic first, and then consider heterogeneous traffic, with and without CBQ.

4.3.1 A single-hop wireless path

For the wireless link, a finite state Markov chain is used to model the time-varying service rates. Let C_j denote the data rate (service rate of the wireless link) in state j, and assume that the total number of states of the wireless link is N_l. We then have

$\mathbf{C} = \mathrm{diag}\{\mathrm{C}_1, \mathrm{C}_2, ..., \mathrm{C}_{\mathrm{N_l}}\}$. The generating matrix of the underlying continuous time Markov process $\mathbf{B}_c$ is defined similar to (4.2) with the dimension of N_l. The link and the incoming traffic are two independent Markovain processes with number of states, N_l and N_s, respectively. The aggregated system can be represented by a Markov chain with number of states, $N_l \times N_s$. It was proven in [30] that if the coupled source and link system states $s = (i, j)$ are ordered lexicographically

$$(1,1)\ (1,2)\ \cdots\ (1,N_l)\ \cdots\ (N_s,N_l),$$

the number of states aggregated system k is equal to $N_l(i-1)+j$ for $1 \leq k \leq N_l N_s$. The generating matrix of the integrated chain can then be obtained as

$$\mathbf{B} = \mathbf{B}_s \oplus \mathbf{B}_c \triangleq \mathbf{B}_s \otimes \mathbf{I}_{N_l} + \mathbf{I}_{N_s} \otimes \mathbf{B}_c. \tag{4.6}$$

Here, $\oplus$ and $\otimes$ are the Kronecker sum and Kronecker product [31], respectively. In (4.6), $\mathbf{I}_{N_l}$ is an $N_l \times N_l$ identity matrix, thus $\mathbf{B}$ is a matrix with dimension $N_l N_s$. $\mathbf{D}$ is the diagonal matrix

$$\mathbf{D} = \mathbf{R}_s \oplus [-\mathbf{C}], \tag{4.7}$$

where $\mathbf{R}_s = \mathrm{diag}\{\mathrm{R}_1, \mathrm{R}_2, ..., \mathrm{R}_{\mathrm{N_s}}\}$. Given $\mathbf{B}$ and $\mathbf{D}$, the probability of buffer overflow can be obtained by solving (4.1). From the perspective of the bottleneck buffer, the output data rate can be viewed as a deduction on the instantaneous incoming traffic rate in each state.

The effective capacity [32] reflects the *stochastic bounded capacity* for a (or a number of) traffic source(s) accessing the buffer, conditioned on the loss probability

P_L and the maximum buffer size x [33]

$$C_e = \frac{1}{\theta} z(\mathbf{B} + \theta \mathbf{D}),$$

where

$$\theta = \frac{1}{x} ln(1/P_L),$$

and $z(\mathbf{A})$ represents the largest eigenvalue of matrix $\mathbf{A}$. High variance input traffic has a higher effective capacity, which leads to fewer connections accommodated by a link. On the other hand, with highly variable output links, the effective capacity will decrease, which also results in fewer admitted connections. This tendency is also demonstrated by the simulation results presented in the next chapter.

4.3.2 A multi-hop wireless path

In this section, the fluid model is further extended to multi-hop wireless paths. Assume that the indoor multi-hop wireless path consists of k_h hops. These k_h hops are within the interference range of each other, so they cannot transmit simultaneously. This is a reasonable assumption for a home networking environment. Since different hops require separate channel times in IEEE 802.15.3 CTAPs, the transmission rate of a packet over a k_h-hop path is

$$C = \left[\sum_{m=1}^{k_h} 1/C_m \right]^{-1}, \tag{4.8}$$

where C_m is the link service rate of the m-th hop.

In the following, we use a 3-hop wireless path as an example. Let $\mathbf{B}_c^{(1)}$, $\mathbf{B}_c^{(2)}$ and $\mathbf{B}_c^{(3)}$ be the generating matrices for the underlying continuous time Markov chains for the first, second and third hops, respectively. If each hop has three states correspond-

ing to three different service rates, as shown in Figure 2.2, each of these matrices will be a 3×3 matrix with components represented by μ_{ij}, $1 \leq i \leq 3$, $1 \leq l \leq 3$. Each hop is an independent Markovain process, thus they can be aggregated into a new system using the same approach as in the last section. Let the states of the multi-hop path be ordered lexicographically. Thus, the generating matrix for the 3-hop path, $\mathbf{B}_c$, has dimension 27

$$\mathbf{B}_c = \mathbf{B}_c^{(1)} \oplus \mathbf{B}_c^{(2)} \oplus \mathbf{B}_c^{(3)}.$$

In each state, the corresponding service rate can be obtained using (4.8). The diagonal matrix D is obtained from (4.7). $\mathbf{C}$ has dimension 27 and C_i is the end-to-end data rate of the 3-hop path for state i.

After obtaining $\mathbf{B}$ and $\mathbf{D}$ according to (4.4), with (4.5), the probability of buffer overflow in the multi-hop case, $G(x)$, can be obtained. Thus, given the QoS requirements of IPTV traffic, the admission region and buffer size for the multi-hop path can be derived.

4.4 Case Study III: Heterogeneous Traffic

In addition to IPTV, some links like those in home networks also accommodate other applications, e.g. data and voice. For the queue system, the previous sections have considered variations in the output process (transmission links). Here we further consider heterogeneous input traffic, and extend the fluid flow analytical framework to consider heterogeneous traffic with and without CBQ.

4.4.1 Without CBQ

First, we assume that all the input traffic sources are independent and multiplexed into a single queue. The burstiness of the data traffic and the regularity of the voice

traffic can be captured by an irreducible, reversible Markov process [34] [35]. Denote $\mathbf{B}_s^{(k)}$, $0 \leq k \leq K_s$, as the infinitesimal generator matrix for the underlying Markov process of the k-th source with $N^{(k)}$ states. Then, the generating matrix representing the combination of input traffic can be obtained by

$$\mathbf{B}_s = \mathbf{B}_s^{(1)} \oplus \mathbf{B}_s^{(2)} \cdots \oplus \mathbf{B}_s^{(K_s)}, \tag{4.9}$$

where the indices of $\mathbf{B}_s$ are ordered lexicographically. The traffic arrival rate $\mathbf{R}_s$ is composed of the superposition of arrival rates for each traffic source

$$\mathbf{R}_s = \mathbf{R}_s^{(1)} \oplus \mathbf{R}_s^{(2)} \cdots \oplus \mathbf{R}_s^{(K_s)}. \tag{4.10}$$

Then the diagonal matrix is represented as $\mathbf{D} = \mathbf{R}_s - C\mathbf{I}$ where C is the constant link service rate.

If the output link is a time-varying single hop (or multi-hop) wireless link, $\mathbf{B}_c$ is the corresponding generating matrix. Matrix $\mathbf{B}$ can be represented as

$$\mathbf{B} = \mathbf{B}_s \oplus \mathbf{B}_c = \mathbf{B}_s^{(1)} \oplus \mathbf{B}_s^{(2)} \cdots \oplus \mathbf{B}_s^{(K)} \oplus \mathbf{B}_c, \tag{4.11}$$

which is obtained by substituting (4.9) into (4.6). $\mathbf{D}$ is given by (4.7). Here, $\mathbf{R}_s$ should be calculated using (4.10). A similar substitution is needed when heterogeneous traffic is relayed by multi-hop wireless networks. Once we obtain $\mathbf{B}$ and $\mathbf{D}$, the probability of buffer overflow can be obtained when heterogeneous traffic is multiplexed.

Remarks: To obtain the buffer overflow probability of heterogeneous traffic multiplexed into constant capacity wired links, time varying single-hop links, or multi-hop wireless links, a similar approach is used. The key is to obtain the generating matrix

of the *aggregated* system, and derive **B** and **D** accordingly. Variations in the traffic rate and channel bandwidth have opposite effects on the admission region. The input rate of the data traffic, for example, from the IPTV traffic point of view, can be considered as a deduction from the available capacity of the output link. In addition, variation in the channel service rate (the deduction from the maximum channel capacity), can be seen as a variable input data rate. This observation matches the theory of effective capacity discussed in Section 4.3.1.

4.4.2 With CBQ

To guarantee the stringent QoS requirements of IPTV, while at the same time protecting data traffic from starvation, CBQ can be employed. There are two virtual queues in the system: one for IPTV traffic and another to accommodate data traffic. Each traffic class can only consume its assigned bandwidth if the other virtual queue is not empty; otherwise, it can use all of the available link bandwidth. Assume that a fraction $(1-p)$ of the service rate is assigned to IPTV traffic and the remainder p is assigned to data traffic when both queues are not empty [1].

Denote the queue occupancy of video and data as F_v and F_d, respectively. The evolution equations of F_v and F_d are

$$\begin{aligned}
\frac{dF_v}{dt} &= \gamma^{(v)}(s_v) + pC - C \qquad F_d > 0, \\
&= \gamma^{(v)}(s_v) + \gamma^{(d)}(s_d) - C \qquad F_d = 0, \\
\frac{dF_d}{dt} &= \gamma^{(d)}(s_d) + (1-p)C - C \qquad F_v > 0, \\
&= \gamma^{(d)}(s_d) + \gamma^{(v)}(s_v) - C \qquad F_v = 0,
\end{aligned} \tag{4.12}$$

[1] Although we are assuming two classes of traffic, the analysis presented here can easily be extended to any number of classes.

where the arrival rates of video and data are $\gamma^{(v)}(s_v)$ and $\gamma^{(d)}(s_d)$, respectively. $\mathbf{S} = (s_v, s_d)$ denotes the combined system state. The queue distribution derivations for video and data are similar. We take the video queue as an example.

The matrix $\mathbf{B}_v$ for the IPTV traffic is given by

$$\mathbf{B}_v = \mathbf{B}_s^{(v)} \oplus \mathbf{B}_s^{(d)},$$

where $\mathbf{B}_s^{(v)}$ and $\mathbf{B}_s^{(d)}$ are the generating matrices of the underlying Markov chains for video and data traffic, respectively. Although $\mathbf{B}_d$ and $\mathbf{B}_v$ may be equivalent for the video and data queue, $\mathbf{D}_d$ and $\mathbf{D}_v$ are different. $\mathbf{D}_v$ is obtained as

$$\mathbf{D}_v = \mathbf{R}_s - C\mathbf{I} = \mathbf{R}_v \oplus \mathbf{R}_{d\prime} \oplus \mathbf{R}_c - C\mathbf{I}, \tag{4.13}$$

where $\mathbf{R}_v$ is the video arrival rate $\gamma^{(v)}(s^{(v)})$. When the arrival rate of the data traffic $\gamma^{(d)}(s_d)$ is larger than its assigned bandwidth pC, the link-sharing scheduler in CBQ will limit the data traffic to its assigned bandwidth. Thus, at those states $s = (s_v, s_d)$, $\gamma^{(d)}(s_d) > pC$, the arrival rate of the data is pC instead of $\gamma^{(d)}(s_d)$ from the video traffic point of view. $\mathbf{R}_{d\prime} = \min(pC, \gamma^{(d)}(s_d))$ is the arrival rate for video. The coefficients a_i can be obtained from the boundary conditions, i.e., $F_o(0) = 0$ for $R(s) > C$. Finally, we can obtain the packet loss ratio of the video packets.

From (4.12), buffer overflow can only occur in those states where $F_v > 0$ (when the overall traffic arrival rate is greater than the total service rate). In equilibrium, when the queue is not filling, the probability of queue overflow is 0, $F_v = 0$. The combined Markov process can be separated into two *virtual* states: Ω_o with $F_v > 0$

and Ω_u with $F_v = 0$. Then we have

$$\Pr(F_v \geq x) = \Pr(F_v \geq x|\Omega_o)\Pr(\Omega_o).$$

In state Ω_o, F_v is filling and video uses up all its reserved bandwidth. Overflow only happens in those states where the total arrival rate is larger than the total capacity. In this case, congestion will occur and CBQ will limit the over-limit traffic to its allocated bandwidth.

Given the tolerable loss rates of arriving traffic, we can determine the admission region of the IPTV service. At the same time, we can appropriately apportion bandwidth for IPTV and data traffic. For the multi-hop wireless case, we use the same approach as previous and obtain B and D according to (4.2) and (4.7).

4.5 Summary

In this chapter, a fluid flow based analytical framework is given. The queue distribution is derived from a set of derivative equations, given the generating matrix, $\mathbf{B}$ and the arrival rate matrix, $\mathbf{D}$. The analysis is used on various network scenarios, considering both traffic and link characteristics. It can help design and plan suitable networks, and choose proper resource allocation schemes for IPTV distribution. In addition, the analysis provides insights into important network parameters or traffic characteristics that would affect network performance. It demonstrates the necessary of traffic shaping for IPTV system. The effectiveness and fidelity of the analysis are verified by extensive NS-2 simulations with real video traces in the next chapter.

Chapter 5

Performance Evaluation by Simulation

To verify the analytical framework and demonstrate the effectiveness of the proposed video model in network simulations and analysis, the analytical results are compared with simulation results with real video traces. Extensive simulations have been performed using the NS-2 [36]. In this chapter, we present the simulation and analytical results for different cases we considered.

5.1 Parameter Settings

The network topology used in the simulations is shown in Figure 5.1. There are N video and/or data source and destination pairs. A source is connected to router $R1$ by a 100 Mbps link. The video stream is forwarded by two routers $R1$ and $R2$ before arriving at the destination. We assume that the video and data traffic is encapsulated in 1000 byte UDP packets. All sinks are connected to router $R2$. The link between $R1$ and $R2$ is the bottleneck link under investigation. It would be a wired link, a single-hop wireless link, or a multi-hop wireless path. The bottleneck link buffer resides in $R1$. The routers use the simple Drop-Tail queue management scheme for solo IPTV traffic, while CBQ is implemented for the heterogeneous traffic.

To simulating IEEE 802.15.3 [37] wireless networks, the data rate of each wireless

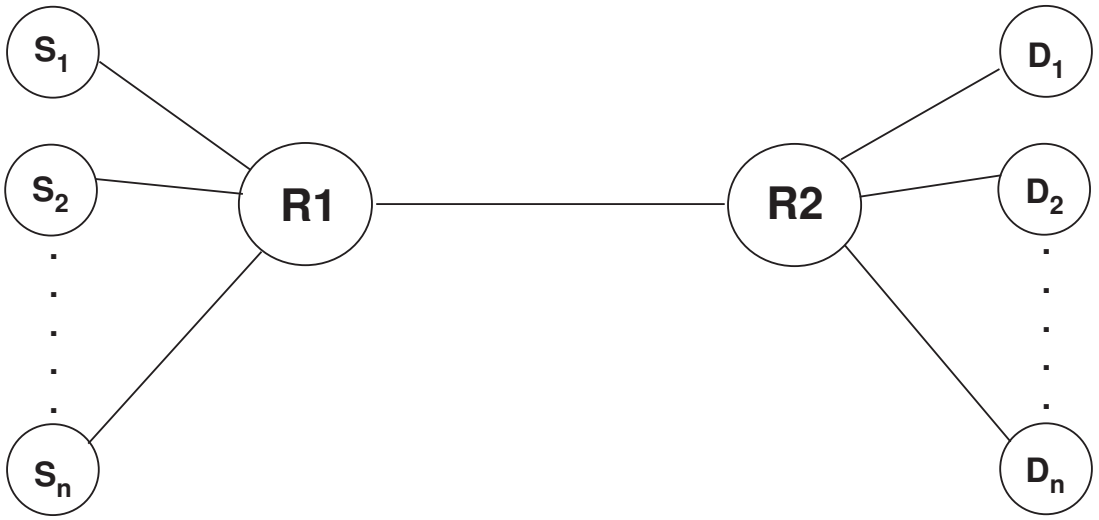

Figure 5.1: The network topology.

link followed a three state Markov chain. The payload data rate is obtained by considering the physical layer data rate and the protocol overheads. The payload transmission time, T_{pay} equals $\frac{payload}{R_i}$, and R_i can be 55 Mbps, 110 Mbps, or 200 Mbps, depending on the current channel state. The time to transmit a frame, T_{frame}, is

$$T_{frame} = T_a + (\frac{payload + RUI_h}{R_i}) + (\frac{PHY_h + MAC_h + HCS + FCS}{R_1}),$$

where $T_a = 15$ μs, $RUI_h = 40$ Bytes, $PHY_h = 2$ Bytes, $MAC_h = 10$ Bytes, $HCS = 2$ Bytes and $FCS = 4$ Bytes are the preamble time, RTP/UDP/IP headers, PHY header, MAC header, Header Check Sequence and Frame Check Sequence, respectively. The PHY/MAC overheads are transmitted at a base data rate R_1 of 55 Mbps. The achievable throughput for the payload is then given as

$$C_i = \frac{T_{pay} R_i}{T_{frame} + T_g + T_{ACK} + 2SIFS},$$

where the guard time $T_g = 10$ μs, the ACK time $T_{ACK} = T_a + (\frac{PHY_h + MAC_h + HCS}{R_1})$,

the SIFS time equal 10 μs and all other overheads take the same values as in [38]. This leads to service rates of $C_1 = 40.3$ Mbps, $C_2 = 64.0$ Mbps and $C_3 = 88.7$ Mbps in states s_1, s_2 and s_3, respectively.

The state transition probability of the wireless link are given in the matrix $\mathbf{P}_c$ below, where P_{ij} in the i^{th} row and j^{th} column is the transition probability from state i to state j.

$$\mathbf{P}_c = \begin{pmatrix} 0.1 & 0.9 & 0 \\ 0.1 & 0.2 & 0.7 \\ 0 & 0.1 & 0.9 \end{pmatrix}.$$

Hence, the state transition rate matrix $\mathbf{B}_c$ is obtained by

$$\mathbf{B}_c = (\mathbf{P}_c - \mathbf{I})/T,$$

where every element of $\boldsymbol{B}_c$, μ_{ij} in the i^{th} row and j^{th} column, is the transition rate from state i to state j, $\mathbf{I}$ is an identity matrix and a wireless channel is changing every T second. The T is set to 20 ms in the case only considering IPTV traffic and 50 ms for the heterogeneous traffic.

Two HD video streams were chosen: "From Mars to China" ("Mars") and "Sony Digital HD Video Camera Demo" ("Sony"). Both are H.264 coded with 1920×1080 resolution. The first is coded from an uncompressed raw video stream, while the latter comes from a WM9 format video file. "Mars" has quantization parameters 28, 28 and 30 for the I, P and B frames, respectively, while the corresponding values for "Sony" are 22, 22 and 24. The larger quantization value, the higher the compression ratio is. With smaller quantization parameters, "Sony" is coded with more details, e.g., its mean I, P and B frame size are around 93.5 KB, 37.5 KB and 10.5 KB,

respectively. The average data rate of "Sony" is 5.8 Mbps, which is higher than that of "Mars" with 4.8 Mbps. The peak data rate of "Sony", however, is lower than "Mars": 50.6 Mbps vs. 78.46 Mbps. Hence, "Mars" is a more bursty video stream. The ratio of peak rate to average rate is about 16 for "Mars", which is twice of that of "Sony". We repeat the simulations with different random seeds and each simulation runs for 30 minutes. The simulation results presented are the averages of 20 runs. The minimum buffer size of simulation is 10 packets.

5.2 Case Study I: Wired Backbone Network

For a bottleneck wired link scenario, first, single IPTV flow case is considered to validate the effectiveness of the proposed video source model on estimating network performance. The performance of the proposed model is compared with that of the classic mini-source model. Then multiple IPTV flows are evaluated with two video traces. The comparison on the two video models shows a trade off in terms of its accuracy and scalability.

5.2.1 Single IPTV flow cases

Figure 5.3 and Figure 5.4 show the PLR of one video stream, "Mars", over a 15 Mbps and a 20 Mbps bottleneck link, respectively. The analytical results with the proposed model are plotted as the "Analysis (mean)" curve, which fits well with the simulation results with the real video trace.

Remarks: According to the queueing theory and previous results [19], instead of using the mean process (i.e., using the average of the frame sizes in each state to be the frame size of that state), the inverse mean process (i.e., using $1/\overline{r} = k/\sum_{i=1}^{k} r_i$ to be the frame size of that state) can be used to accurately estimate the packet loss rate in overload states. The analytical and simulation results show that using the simpler

average method (the curve corresponding to "Analysis (Mean)") can achieve similar accuracy in estimating the PLR as the inverse mean method (the curve corresponding to "1/Mean").

To partition the GoP-level states, instead of using our approach which divides the states such that all states have the same steady state probability, two other approaches were also examined. These results are also shown in the figure. The first approach ("Overload state") is to group all I frames with an arrival rate exceeding the link capacity as a single overload state, and group the other frames into two states with the same steady state probability. The second approach ("(Max.-Min.)/3") is to divide the range of I frame sizes into three even states. These two approaches both over-estimate the PLR of video significantly, so it is not advisable to use them. The analytical results with the mini-source model ("Mini-source") are also plotted in the figure. As anticipated, it under-estimates the PLR. Since in the mini-source model, 2.19 Mbps traffic is generated in the "on" state, and we use 8 mini sources to model a video stream. It results in the maximum rate of the input stream is only 17.52 Mbps, while with real trace, the maximum value is as high as 78.46 Mbps. The gap between the mini-source model and simulation results is due to its peak rate not successfully captured by the mini-source model. The results of one "Mars" indicate that 20M link bandwidth is needed to accommodate one single "Mars" connection.

5.2.2 Multiple IPTV flow cases

Figure 5.5 and Figure 5.6 show the performance of multiple video sources of "Mars" and "Sony" sharing an 85 Mbps link, respectively. We use the two-level traffic model to generate synthetic video traces in the simulations, and compare the results with the simulations using the real video traces. As shown in the figure, the resultant PLRs match very well. Thus, the two-level traffic model is an effective tool for fast

simulation purposes. The analytical results using the mini source model can still be used to quantify the performance and the admission region when the buffer size is large. The results show that, to ensure PLR less than 10^{-4}, upto ten "Mars" or eight "Sony" videos could be supported by an 85 Mbps link with a 600 packet buffer.

The results of mini-source model appears to be under-estimate the PLR with relatively small buffer size, and it is close to simulation results as the buffer size increasing. Notice that the accuracy of our model is at the cost of scalability. The number of state for one video trace is 108 for one video stream, which is significant. Nevertheless, the matrix is sparse and simple to handle mathematically in terms of calculating eigenvalues and eigenvectors. If there are n streams multiplexing in the buffer, the integrated mini-source Markov chain has the order of $8n + 1$. However with the proposed two-level Markov model considering the intra-GoP correlation, due to the fully connected Markocv chain, the order is up to 108^n. Hence the mini-source model is more scalable but less accuracy compared with the proposed model. In the remaining chapter, "Mars" is chosen to represent the bursty IPTV traffic.

5.3 Case Study II: Wireless Network

For an one hop wireless link with an average data rate of 85 Mbps, when a buffer size is greater than 200 packets, the PLR drops below 10^{-4} when admitting 6 video streams, while 400 packets is needed to admit 7 videos, as shown in Fig. 5.7. Comparing these results with those for the 85 Mbps wired link, the time-varying wireless link can admit far fewer video streams. This demonstrates that if we simply use the average data rate of a time-varying link to calculate the admission region, the results will be too optimistic.

The results for a three-hop wireless relay path are given in Fig. 5.8. The average

data rate of each hop is still 85 Mbps. To maintain a PLR of 10^{-4}, the maximum number of video streams that can be admitted is now only 2. As anticipated, an increase in the number of hops greatly affects the admission region. In fact, the admission region for the three-hop wireless network is one-third that of the one-hop wireless case and one-fifth that of the 85 Mbps wired link.

For both cases, the mini-source model obtains the accurate admission region but moderate to give the necessary buffer size. It may bring problems in the wireless environment when buffer is considered as a limited resource need to be thoroughly allocated at a mobile node (user). In addition, if an application has a stringent delay-jitter budget, queue size affecting queueing delay is a crucial factor on this issue. However, we argue that for IPTV services, the delay-jitter could be solved by pre-buffering a certain duration of programme in the local set-top box. Normally, the buffer resource is ample for WPAN environment. Thus, the mini-source model is still sufficient to provide the admission region of IPTV traffic.

5.4 Case Study III: Heterogeneous Traffic

To determine the admission region of IPTV connections with existence of data traffic, performance of video traffic with 5 Mbps CBR data traffic and bursty On-Off data traffic with a peak rate 5 Mbps are investigated. For the On-Off data stream, the Off period is 10 times the On period. The average On period is set to be 55.605 sec according to the measurements in [39]. For simplicity, both video and data traffic are encapsulated in 1000 byte packets. A "virtual" CBQ is deployed at the router R1, as shown in Figure 5.2. There are separated queues for different traffic classes, video buffer ranges from 400 packets to 1100 packets, and the data buffer size is set to be the closest integer of $(1/p - 1)$ times the video buffer size. The "virtual" buffer size

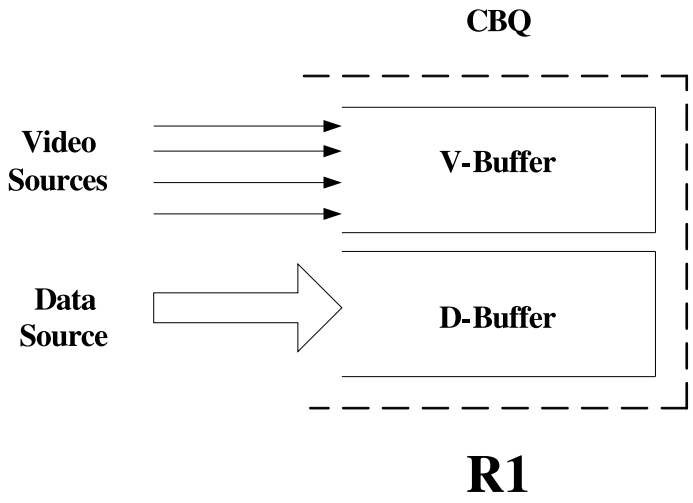

Figure 5.2: CBQ resides at the router R1.

shown in the horizontal axis of Figs. 5.9-5.12 denotes the total buffer for both video and data traffic. Since the buffer size is large, the classic mini-source model is used for analytical purpose, which is still accurate to obtain admission region of bursty IPTV traffic.

5.4.1 Wired backbone network

Figure 5.9 and Figure 5.10 show the analytical and simulation results for buffer overflow probability vs. buffer size, with a CBR and an On-Off data stream, respectively. For the CBQ, 5% of the link bandwidth is reserved for data, and the remaining bandwidth is used to support IPTV traffic. The fluid model provides a good approximation when the buffer size is large, and the admission region derived by the analysis is accurate.

In Figure 5.9, both the analytical and simulation results show that, if one data source keeps transferring at a CBR of 5 Mbps, to ensure a packet loss rate (PLR) less than 10^{-4}, a maximum of 10 IPTV connections can barely be supported with a buffer size larger than 1000 packets. Nine IPTV connections can be supported with

a safe margin.

In a typical home network, data traffic tends to be more bursty than CBR. With typical bursty data traffic, the duration of Off periods can be up to 10 times of the On periods. Given the appropriate link-sharing mechanism, IPTV and data traffic can be efficiently distributed inside the home simultaneously. As shown in Figure 5.10, an 85 Mbps wired link can support 10 IPTV connections and one On-Off data connection. Simulation results (not illustrated) also show that a buffer size of 11 packets is sufficient to guarantee the PLR of data traffic to be less than 10^{-4}. To traverse 11 IPTV connections in the link is not appropriate since the PLR is always above 10^{-4} even with buffer size over 1100 packets.

We conclude that for an 85 Mbps wired link with CBQ, the admission region is 10 video streams in the presence of a 5 Mbps bursty data traffic. In the worst case where the data traffic is a 5 Mbps CBR, the admission region drops to 9 IPTV connections.

5.4.2 Wireless network

The performance of IPTV and CBR data traffic in a 3-hop wireless relay path is shown in Figs. 5.11 and 5.12, respectively. With 5% of the bandwidth reserved for data, it is impossible to support more than one IPTV connection. The PLR of the two traffic classes is up to 10^{-4} for video and 10^{-2} for data when 2 IPTV connections and 1 data connection are sharing the network. Only one IPTV connection could be supported over the 3-hop wireless path.

To study the effect of CBQ on the admission region, we simulated the same scenario, omitting the use of CBQ. We notice that the video performance is degraded, while at the same time, the PLR of the data is lower (i.e., achieving better performance). This behavior was expected because CBQ was used to protect the IPTV traffic from the more aggressive data traffic.

When there is only one IPTV connection over the 3-hop wireless path, the quality of video is ensured, but the PLR of the data traffic is up to 10^{-4}, as shown in Fig. 5.12. If we further increase the buffer size for data, the data PLR would be lower, but this may increase the queueing delay and jitter. Another means of lowering the data PLR involves increasing the reservation for data traffic using CBQ. For the 3-hop wireless path case, if 10% of the total bandwidth is reserved for data traffic, the PLRs for both the IPTV connection and the data connection are below 10^{-6}. Therefore, if bandwidth is reserved appropriately, the IPTV and data performance can be improved, and the QoS for both are guaranteed.

In summary, if we allocate resources for different traffic classes appropriately, both video and data traffic can be supported over a multi-hop wireless path with satisfactory QoS. Although the average throughput of each hop is 85 Mbps, the same as that of the wired link, the admission region for the 3-hop wireless network is only one-ninth that of the wired link. This is because, first, each packet must be transmitted in three time slots to avoid collisions; and second, the time variation of the wireless channel decreases the effective capacity compared to the wired link.

5.5 Summary

The admission regions of the IPTV traffic are summarized in Table 5.1 and Table 5.2. It also gives the minimum buffer size needed to guarantee a PLR less than 10^{-4}. For heterogeneous wireless and wired networks, the admission region is determined by the admission region of the bottleneck link. For the admission regions obtained over the 20 simulation runs, we count the number of runs with PLR above the threshold and obtain the confidence level of the admission region, as shown in the table.

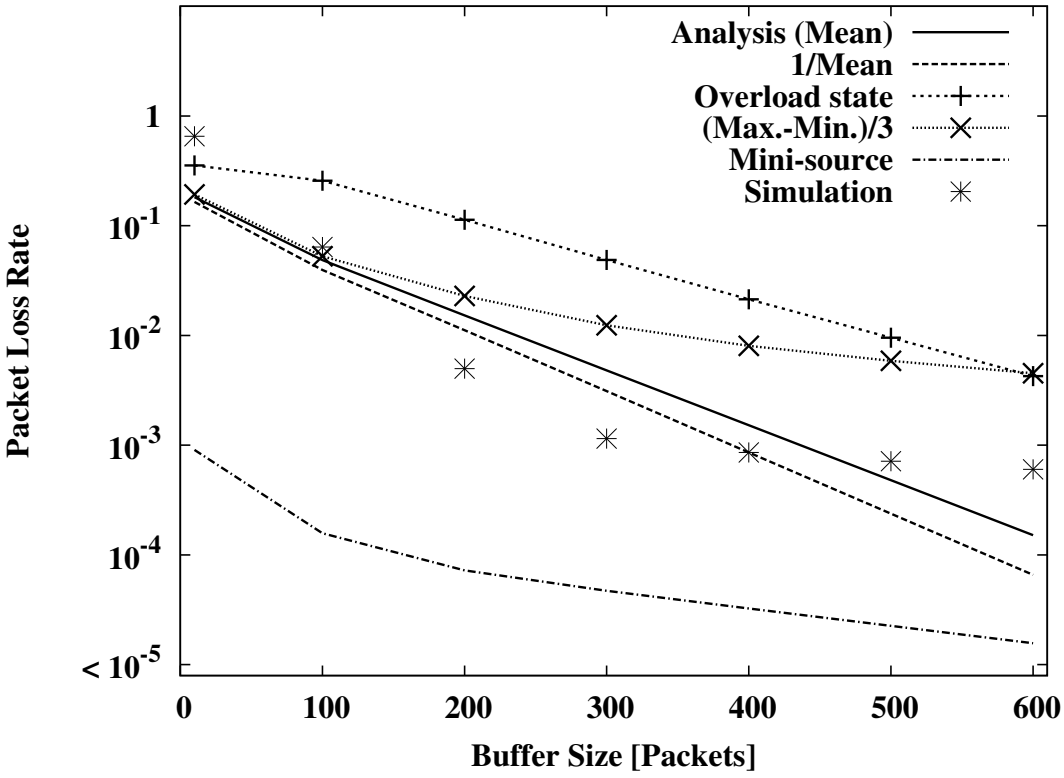

Figure 5.3: PLR of one "Mars" connection with a 15 Mbps link.

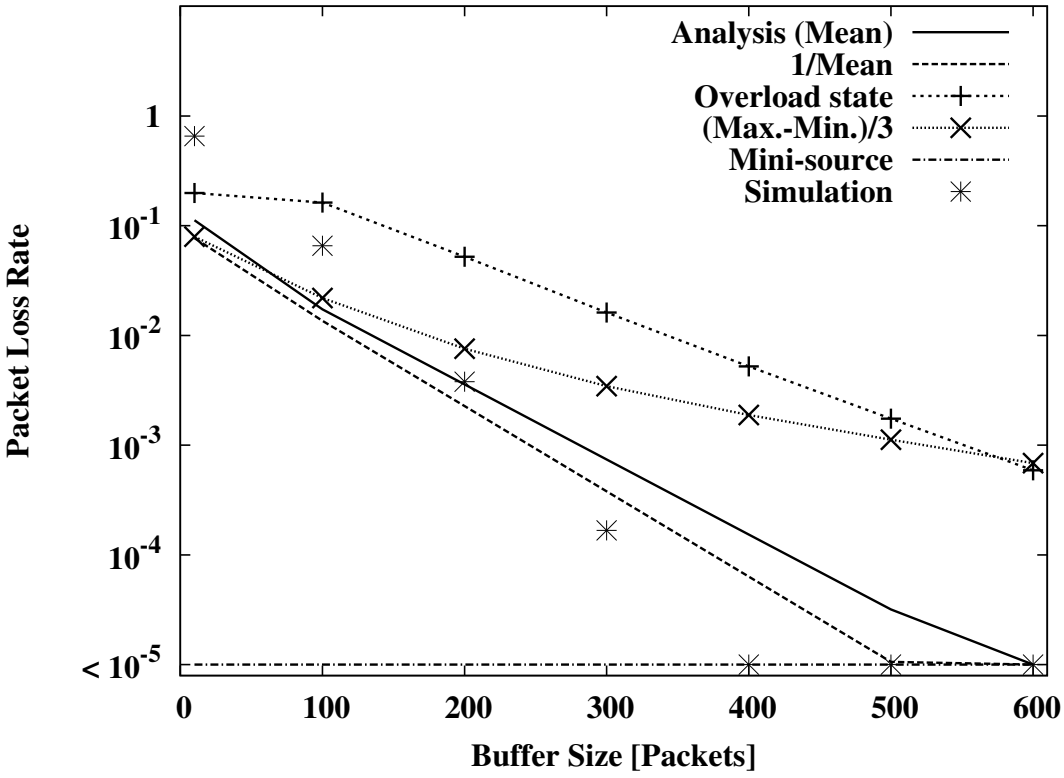

Figure 5.4: PLR of one "Mars" connection with a 20 Mbps link.

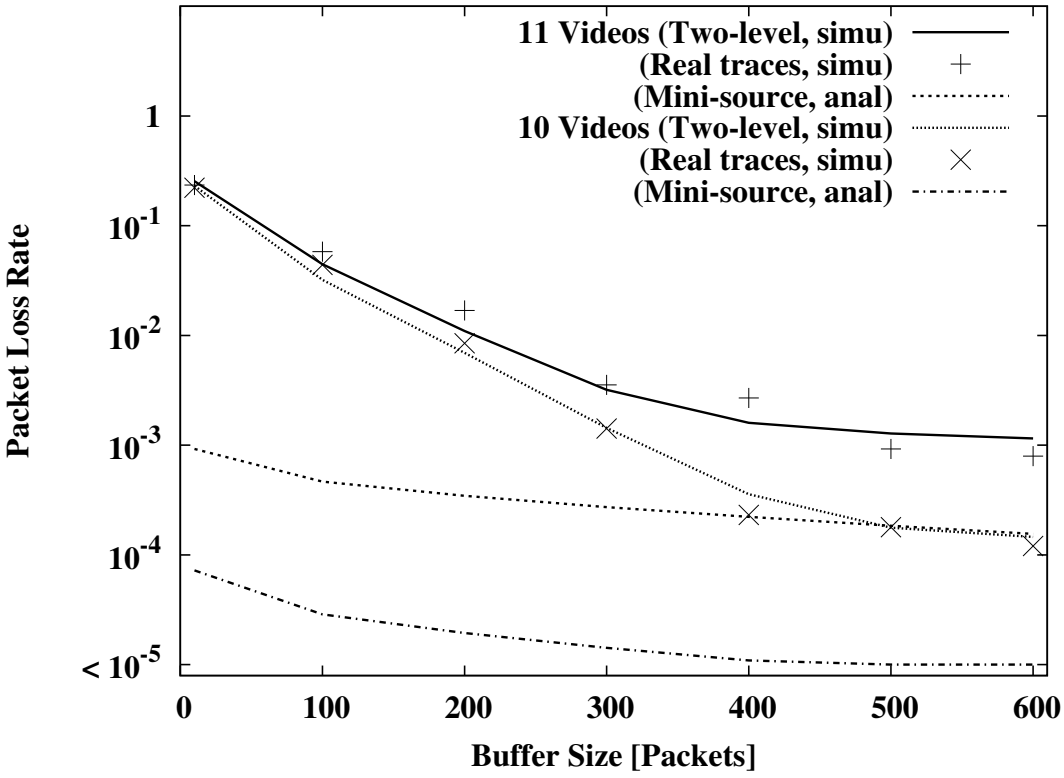

Figure 5.5: PLR of multiple "Mars" connections in a 85Mbps link.

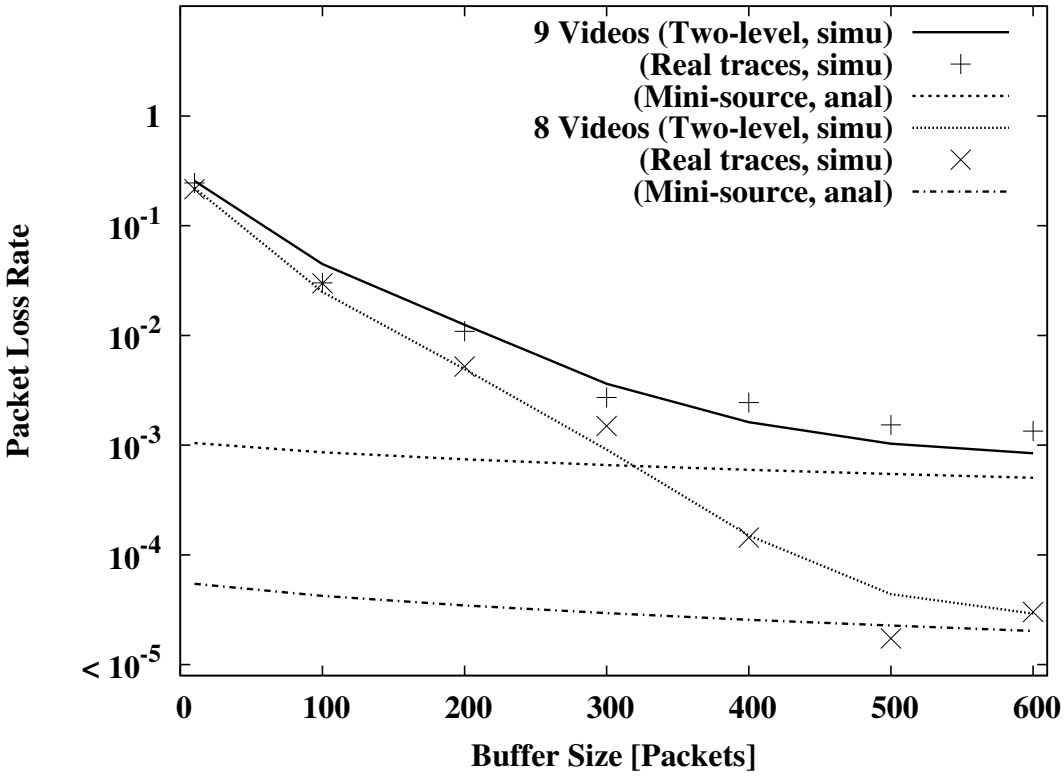

Figure 5.6: PLR of multiple "Sony" connections in a 85Mbps link.

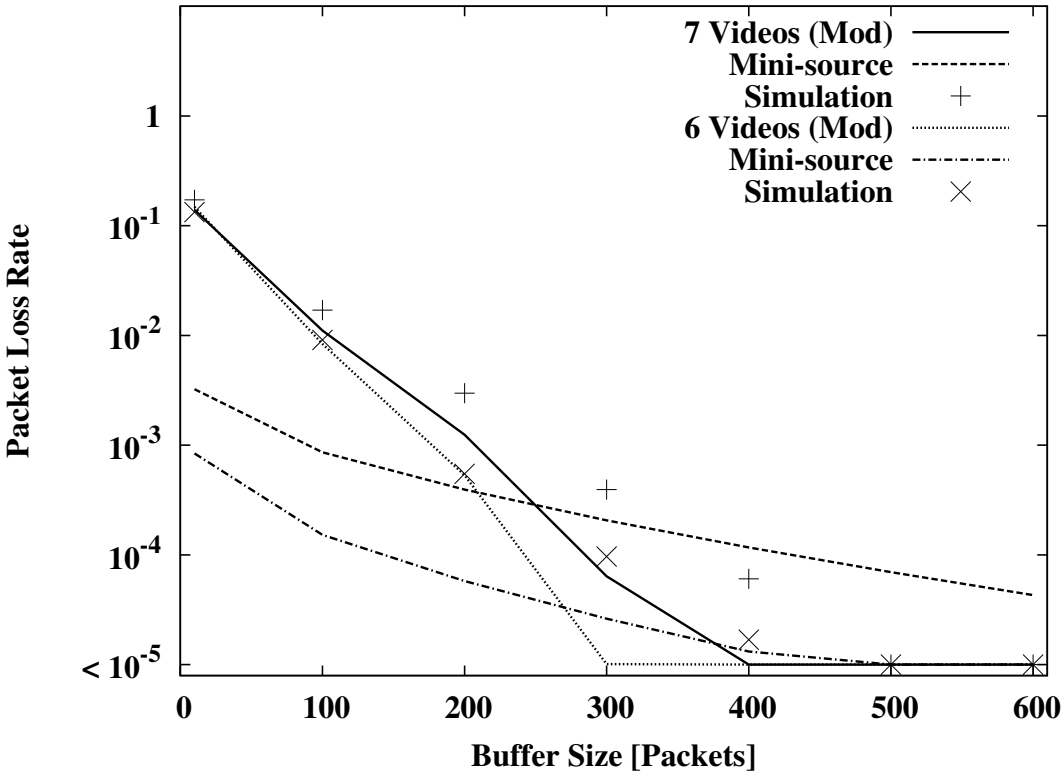

Figure 5.7: PLR of multiple "Mars" connections in a single-hop link.

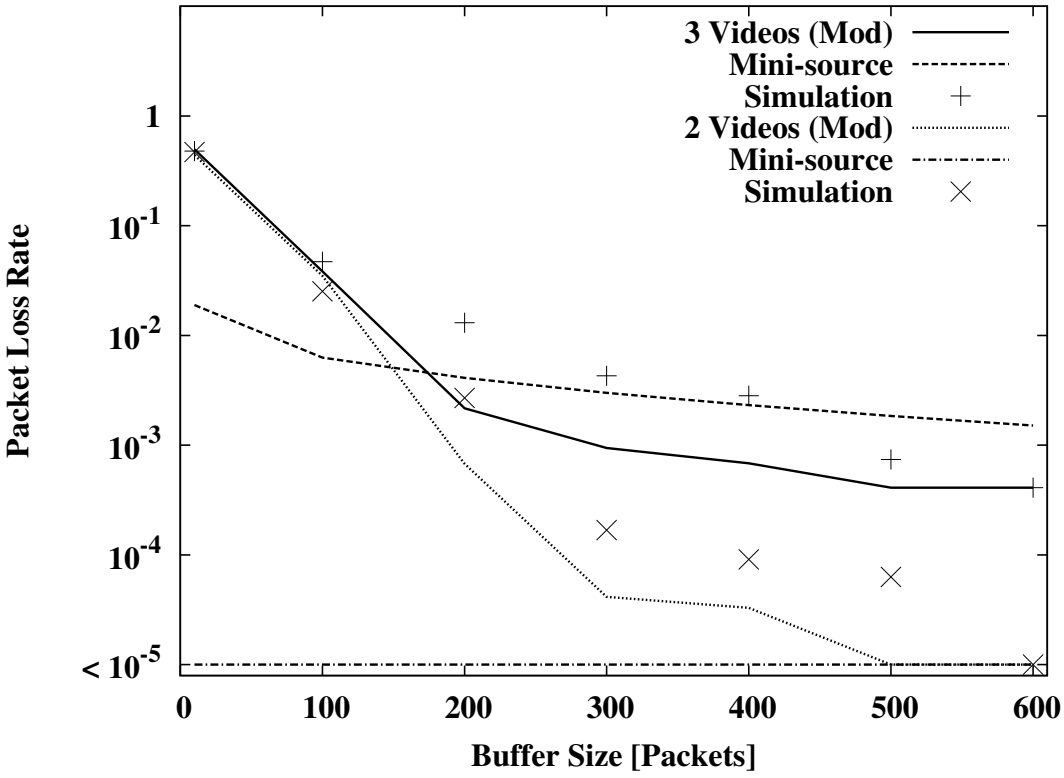

Figure 5.8: PLR of multiple "Mars" connections in a three-hop link.

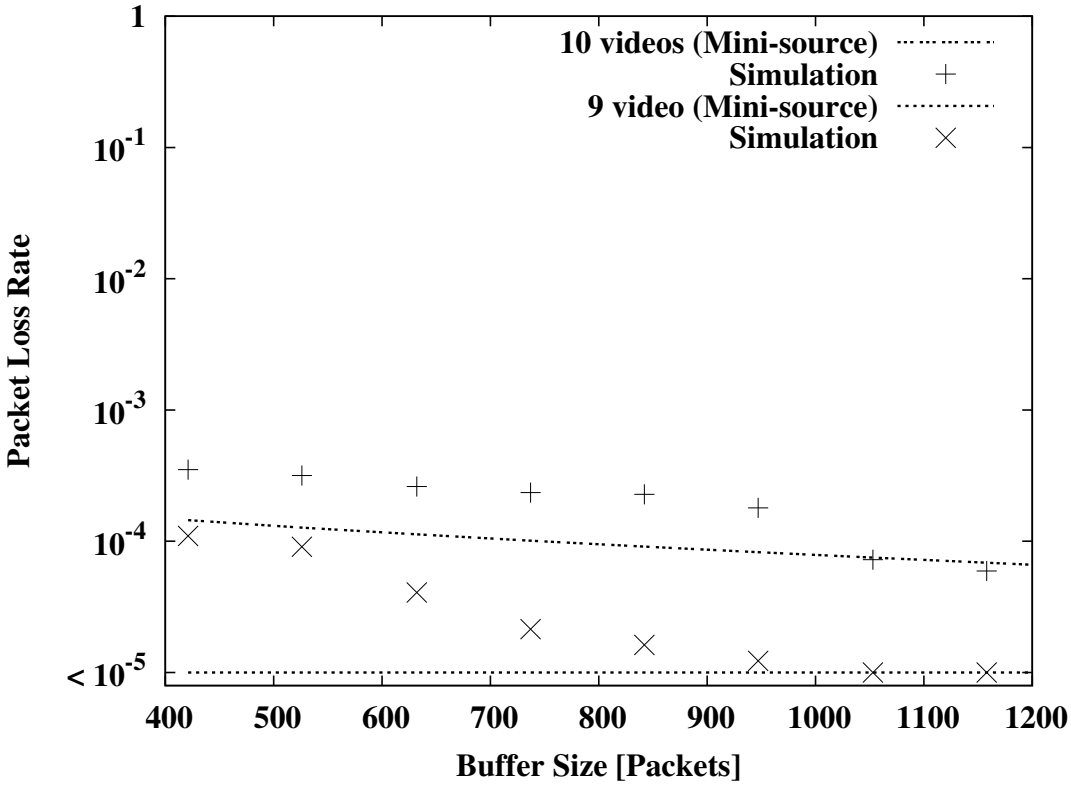

Figure 5.9: Performance of video (with CBR data) in an 85 Mbps wired link.

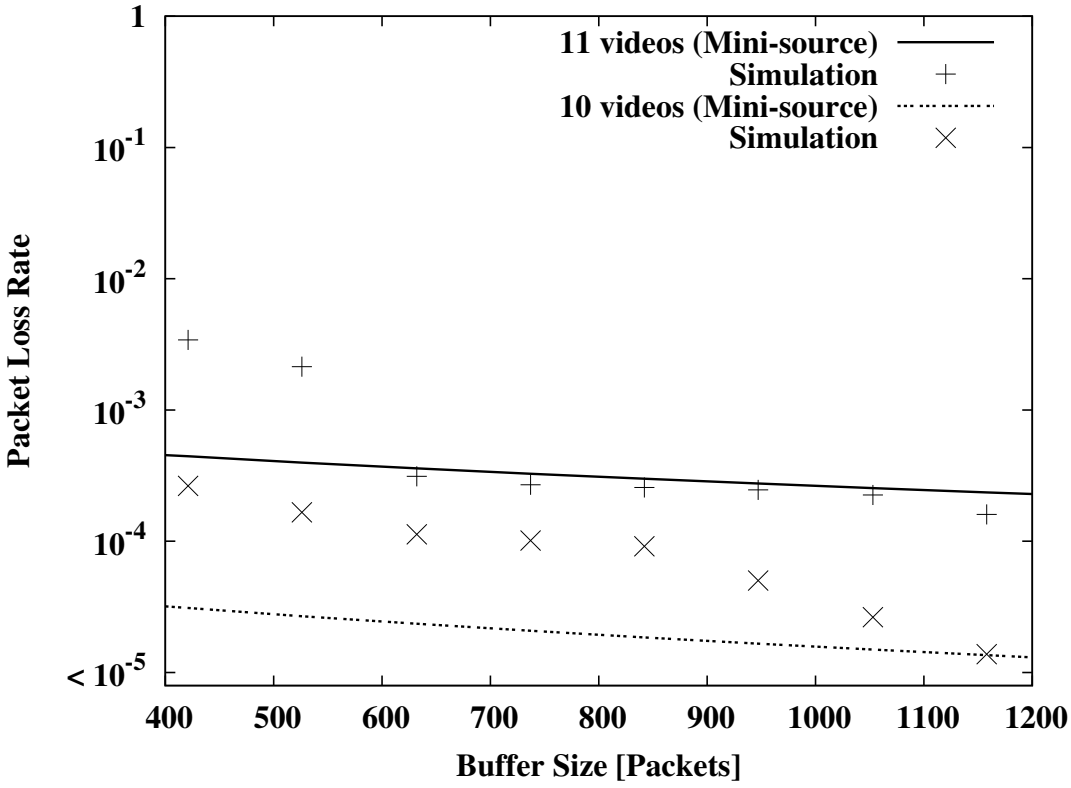

Figure 5.10: Performance of video (with On-Off data) in an 85 Mbps wired link.

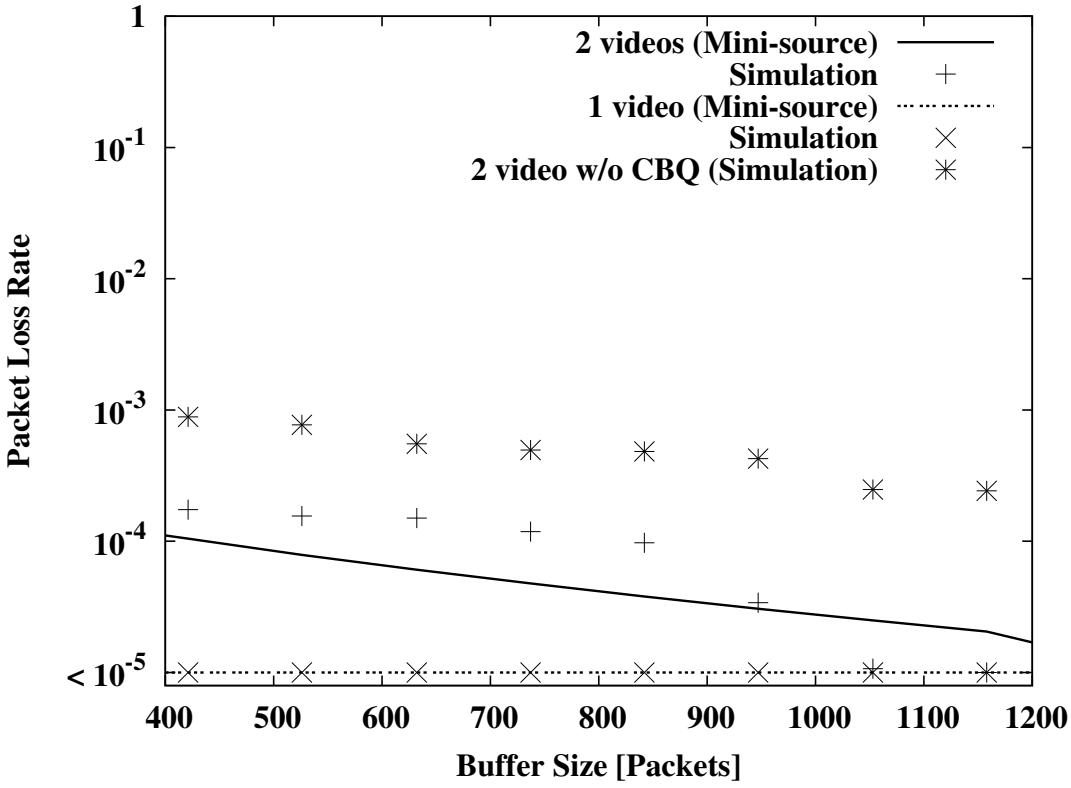

Figure 5.11: Performance of video (with CBR data) in a 3-hop wireless path.

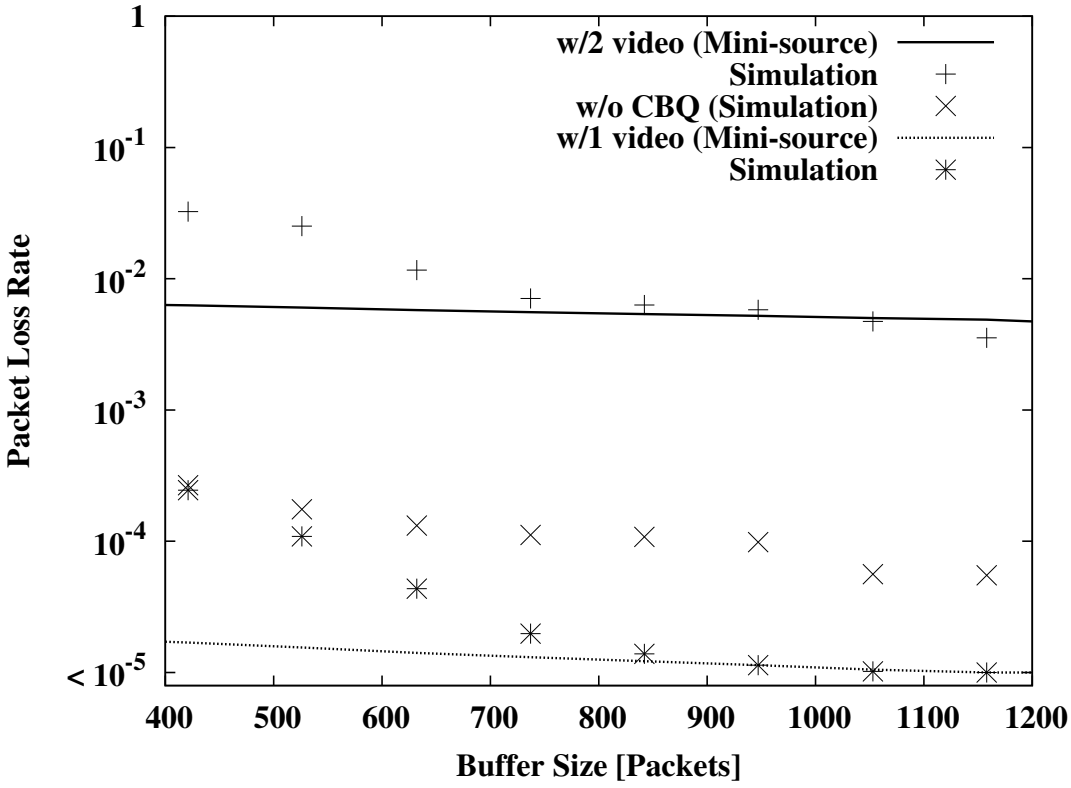

Figure 5.12: Performance of CBR data (with video) in a 3-hop wireless path.

Table 5.1: Admission Region of IPTV Traffic

Link Type	Video Trace	Bandwidth (Mbps)	Admission Region		Buffer Size (Packets)	Confidence Level (%)
			Analysis (Proposed model)	Simulation		
Wired	"Mars"	20	1	1	400	-
	"Mars"	85	10	10	600	85
	"Sony"	85	8	8	500	90
1-hop wireless	"Mars"	-	7	7	400	85
3-hop wireless	"Mars"	-	2	2	400	80

Table 5.2: Admission Region of "Mars" with data traffic

Link Type	Data	Bandwidth Allocated for Data	Admission Region		Buffer Size (Packets)	Confidence Level (%)
			Analysis (Mini-source)	Simulation		
85M Wired	CBR	5%	9	9	400	95
	On-Off	5%	10	10	400	90
3-hop wireless	CBR	5%	1	1	600	90
	CBR	10%	1	1	400	100

Chapter 6

Conclusions and Future Work

6.1 Conclusions and Summary of Contributions

In this thesis, we have investigated the IPTV packet-loss performance due to buffer overflow over wired and multi-hop wireless links, with or without CBQ. In addition, we have quantified the admission region for IPTV streams in a wireless and wired home network, given the link data rate, video traffic statistics, and QoS requirements. In a hybrid network consisting of both wired and wireless links, the admission region is determined by the one with the smallest capacity. The analytical framework can help service providers improve the design and deployment of IPTV systems and provide important insight of which system parameters affect the admission region of networks carrying IPTV streams.

In addition, we have developed a two-level Markovian traffic model for video sources. This model exploits both the spatial and temporal correlations in MPEG encoded video sequences. The bursty HD TV video traces have been investigated to demonstrate the accuracy of the proposed traffic model for studying network performance. The model is simple and can easily be incorporated in network simulators, and can also be used to quantify queue performance analytically. Therefore, the

model can be an effective tool for IPTV performance evaluation via analysis and/or simulations.

The main contributions of this thesis are threefold. First, we propose a simple, two-level Markovian traffic model for IPTV streams to facilitate network performance analysis and simulations. Then, an analytical framework for the queue distribution is built upon the fluid flow approach. The analysis model is adaptable to different network scenarios and traffic types to obtain the admission region of traffic over a bottleneck link. Thirdly, the CBQ is suggested to be deployed for supporting heterogeneous traffic sharing the same network. Both analysis and simulation results indicate its effectiveness for ensuring the QoS of heterogeneous traffic classes. The feasibility of the proposed video model and the accuracy of the analytical results are verified through extensive simulations.

The analytical framework quantifies the impacts of traffic characteristics and channel characteristics on network performance. The main conclusions are as follows. Bursty traffic is more likely to cause buffer overflow, and the more variable wireless link can accommodate less IPTV connections than the wired one with constant data rate. CBQ, as a resource management scheme, bounds the incoming traffic peak rate, thus ensuring QoSs of different traffic classes. From the analysis, it also shows that the instantaneous rate is an important factor that can affect buffer overflow. Based on these observations, we suggest service providers consider the following two aspects in order to deliver IPTV services successfully:

1. A proper resource allocation scheme should be applied for heterogeneous traffic; otherwise, the QoS of all traffic classes might be affected and the network resources cannot be fully utilized.

2. Since smooth traffic is preferred for QoS provisioning, traffic shaping (access

control) should be deployed in the boundary of networks. It will reduce loss rate of traffic due to buffer overflow and reduce queueing delay in the network.

Traffic shaping design and its impacts on IPTV streams are worth for further investigation. Other open issues are listed in the next section.

6.2 Further Research Issues

To ensure the success and QoS of IPTV services, there are many related research issues beckon for further investigation. There are three main open issues that need to be addressed:

1. We assume that all transmissions inside a wireless network are within the same collision domain and through a dedicated path. In a wireless network, if concurrent transmissions, multi-path transmissions, or cooperative transmissions are allowed, how to quantify the queue behavior and maximize the admission region become a complicated problem requires joint efforts from different layers, e.g., how to enhance the wireless MAC protocols to enlarge the admission region will be an interesting further research issue.

2. Proper resource allocation schemes for triple-play services are worth to develop and evaluate in the near future. The scheme should ensure the protection of each traffic class and maximize the performance of network. This issue is even crucial for resource (bandwidth, buffer, power, etc.) limited wireless environment.

3. In addition, the proposed traffic model has been proven to be effective and accurate for network performance analysis and simulations. Its large sparse

generating matrix, however, will be difficult to handle with increasing numbers of sources. The trade-off between the scalability and accuracy has been presented in the thesis. A more scalable solution still needs investigation.

Based on the above conclusions, the proposed video traffic model and the analytical framework for IPTV systems are promising to determine queue distribution and admission region of IPTV connections. Many opportunities for further improvement still exist. We outlined a few key issues in this chapter in the hope of inspiring further research interest in this important area.

Bibliography

[1] V. Jacobson, "Congestion avoidance and control," *SIGCOMM Comput. Commun. Rev.*, vol. 25, no. 1, pp. 157–187, 1995.

[2] D. D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: Architecture and mechanism," in *Proc. ACM SIGCOMM*, Aug. 1992, pp. 14–26.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated service," RFC Editor, United States, RFC2475, 1998.

[4] L. Zhang, S. Deering, and D. Estrin, "RSVP: A new resource ReSerVation protocol," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 116–127, May 2002.

[5] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *SIGCOMM Comput. Commun. Rev.*, vol. 19, no. 4, pp. 1–12, 1989.

[6] J. C. R. Bennett and H. Zhang, "WF2Q: Worst-case fair weighted fair queueing," in *Proc. IEEE INFOCOM*, 1996, pp. 120–128.

[7] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single node case," in *Proc. IEEE INFOCOM*, 1992, pp. 915–924.

[8] A. K. Parekh and R. G. Gallagher, "A generalized processor sharing approach to flow control in integrated services networks: the multiple node case," *IEEE/ACM Trans. Networking*, vol. 2, no. 2, pp. 137–150, 1994.

[9] I. Stoica, S. Shenker, and H. Zhang, "Core-stateless fair queueing: Achieving approximately fair bandwidth allocations in high speed networks," *SIGCOMM Comput. Commun. Rev.*, vol. 28, no. 4, Sept. 1998.

[10] S. Floyd and V. Jacobson, "Link-sharing and resource management models for packet networks," *IEEE/ACM Trans. Networking*, vol. 3, no. 4, pp. 365–386, Aug. 1995.

[11] A. Elwalid and D. Mitra, "Analysis, approximations and admission control of a multi-service multiplexing system with priorities," in *Proc. IEEE INFOCOM*, Apr. 1995, pp. 463–472.

[12] F. Presti, Z. Zhang, and D. Towsley, "Bounds, approximations and applications for a two-queue GPS system," in *Proc. IEEE INFOCOM*, Mar. 1996, pp. 1310–1317.

[13] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE/ACM Trans. Networking*, vol. 36, no. 7, pp. 834–844, July 1988.

[14] E. Shihab, F. Wan, L. Cai, T. A. Gulliver, and N. Tin, "Performance analysis of IPTV traffic in home networks," in *Proc. IEEE GLOBECOM*, Nov. 2007, pp. 5341–5345.

[15] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell System Tech. J.*, pp. 1253–1265, Sept. 1960.

[16] E. Shihab, L. Cai, F. Wan, T. A. Gulliver, and N. Tin, "Wireless mesh networks for in-home IPTV distribution," *IEEE Network*, vol. 22, no. 1, pp. 52–57, Jan./Feb. 2008.

[17] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," *Proc. IEEE Trans. Vehicular Technology*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[18] M. Dai and D. Loguinov, "Analysis and modeling of MPEG-4 and H.264 multilayer video traffic," in *Proc. IEEE INFOCOM*, Mar. 2005, pp. 2257–2267.

[19] P. Skelly, M. Schwartz, and S. Dixit, "A histogram-based model for video traffic behavior in an ATM multiplexer," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 446–459, Aug. 1993.

[20] A. Dawood and M. Ghanbar, "Content-based MPEG video traffic modeling," *IEEE Trans. Multimedia*, vol. 1, no. 1, pp. 77–87, Mar. 1999.

[21] U. K. Sarkar, S. Ramakrishnan, and D. Sarkar, "Modeling full-length video using Markov-modulated gamma-based framework," *IEEE/ACM Trans. Networking*, vol. 11, no. 4, pp. 638–649, Aug. 2003.

[22] B. Melamed and D. Pendarakis, "Modeling full-length VBR video using Markov-renewal-modulated TES models," *IEEE J. Selected Areas in Commun.*, vol. 16, no. 5, pp. 600–611, June 1998.

[23] M. M. Krunz and A. M. Makowski, "Modeling video traffic using M/G/∞ input processes: a compromise between markovian and LRD models," *IEEE J. Selected Areas in Commun.*, vol. 16, no. 5, pp. 733–748, June 1998.

[24] M. Schwartz, *BroadBand Integrated Network*. Prentice Hall, 1996.

[25] D. P. Heyman and T. V. Lakshman, "What are the implication of long-range dependence for VBR-video traffic engineering," *IEEE/ACM Trans. Networking*, vol. 4, no. 3, pp. 301–317, June 1996.

[26] "H.264/MPEG-4 part 10 standard," ITU-T Video Coding Experts Group and ISO/IEC Moving Picture Experts Group, Tech. Rep. [Online]. Available: http://www.itu.int/rec/T-REC-H.264

[27] Tech. Rep. [Online]. Available: http://trace.eas.asu.edu/hd/index.html

[28] A. Lombardo, G. Morabito, and G. Schembra, "An accurate and treatable markov model of MPEG-video traffic," in *Proc. IEEE INFOCOM*, Mar./Apr. 1998, pp. 217–224.

[29] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *Proc. ACM SIGCOMM*, Aug./Sept. 1994, pp. 269–280.

[30] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Advances in Applied Probability*, vol. 20, no. 3, pp. 646–676, Sept. 1988.

[31] A. J. Laub, *Matrix Analysis for Scientists and Engineers*. SIAM Publications, 2005.

[32] C.-S. Chang, "Stability, queue length, and delay, II: Stochastic queueing networks," IBM Research, Yorktown Hts., NY, Report RC 17709, Feb. 1992.

[33] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[34] P. T. Brady, "A model for generating on-off speech patterns in two-way conversations," *Bell Sys. Tech. J.*, vol. 48, pp. 2445–2472, Sept. 1969.

[35] L. Kleinrock, *Queueing Systems, Vol. I: Theory.* John Wiley, 1975.

[36] "The network simulator – ns-2," Tech. Rep. [Online]. Available: http://www.isi.edu/nsnam/ns

[37] "IEEE standard 802 part 15.3: Wireless medium access control (MAC) and physical layer (PHY) specifications for high rate wireless personal area networks (WPAN)," IEEE, Tech. Rep., 2003.

[38] L. X. Cai, L. Cai, X. S. Shen, and J. W. Mark, "Capacity of UWB networks supporting multimedia services," in *Proc. QShine '06*, Aug. 2006.

[39] S. Deng, "Empirical model of WWW document arrivals at access link," in *Proc. IEEE Int. Conf. Commun. (ICC)*, June 1996, pp. 1797–1802.